Reports in Phonetics, University of the Saarland

Berichte zur Phonetik, Universität des Saarlandes

PHONUS

Edited by: W. J. Barry, B. Möbius & J. Trouvain

No. 17, March 2014

Institut für Phonetik
Universität des Saarlandes
D-66041 Saarbrücken, März 2014

ISSN 0949-1791

VORWORT DER HERAUSGEBER

Die PHONUS-Reihe setzt die Veröffentlichung von Doktorarbeiten von Mitgliedern der Phonetik-Gruppe an der Universität des Saarlandes fort. Der vorliegende Band, PHONUS 17, präsentiert Eva Lasarcyks Doktorarbeit mit dem Titel *Empirical evaluation of the articulatory synthesizer VocalTractLab as a discovery tool for phonetic research: Articulatory-acoustic investigations of paralinguistic speech phenomena*. In ihrer Dissertation untersucht Eva Lasarcyk eine Reihe von paralinguistischen Phänomenen der gesprochenen Sprache mit Hilfe eines artikulatorischen Synthesesystems. Die Arbeit verbindet experimentalphonetische Fragestellungen und Methoden mit sprachtechnologischen Werkzeugen. Der derzeit wohl am weitesten ausgereifte artikulatorische Synthetisator, *VocalTractLab* (entwickelt von Peter Birkholz, Rostock/Aachen), wird in den Experimenten hinsichtlich seiner Fähigkeiten der Modellierung artikulatorischer Prozesse voll ausgereizt. Eva Lasarcyk weist nach, dass sich *VocalTractLab* als Werkzeug für die Untersuchung von Details der Sprachproduktion grundsätzlich sehr wohl eignet, allerdings in einigen phonetischen und technischen Aspekten auch an seine Grenzen stößt. Aus phonetischer Sicht ist vor allem die Modellierung der Synchronisierung und Phasierung artikulatorischer Gesten noch verbesserungsbedürftig, und aus technischer Sicht ist die Erweiterung in Richtung auf eine textbasierte Synthese (*text-to-speech synthesis*, TTS) wünschenswert.

Saarbrücken, im März 2014 William J. Barry, Bernd Möbius & Jürgen Trouvain

EDITORS' FOREWORD

The PHONUS series continues to publish doctoral theses by members of the Phonetics group at Saarland University. The current volume, PHONUS 17, presents Eva Lasarcyk's PhD dissertation, entitled *Empirical evaluation of the articulatory synthesizer VocalTractLab as a discovery tool for phonetic research: Articulatory-acoustic investigations of paralinguistic speech phenomena*. In her thesis, Eva Lasarcyk investigates several paralinguistic phenomena in spoken language by means of an articulatory speech synthesizer, combining experimental-phonetic research questions and methods with speech technology tools. *VocalTractLab* (developed by Peter Birkholz, Rostock/Aachen), arguably the most advanced articulatory synthesizer today, is put to the test with respect to its capabilities of modeling articulatory processes. Eva Lasarcyk demonstrates that *VocalTractLab* is in principle a highly useful tool for investigating details of speech production but is also constrained by limitations in certain phonetic and technological aspects. From a phonetic point of view, modeling the synchronization and phasing of articulatory gestures needs further improvement, and from a technological point of view, the extension of *VocalTractLab* to a full-fledged text-to-speech synthesizer would be highly desirable.

Saarbrücken, March 2014 William J. Barry, Bernd Möbius & Jürgen Trouvain

Empirical evaluation of the articulatory synthesizer VocalTractLab as a discovery tool for phonetic research: Articulatory-acoustic investigations of paralinguistic speech phenomena

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Eva Lasarcyk

aus Freiburg im Breisgau

Dekan:	Prof. Dr. Roland Marti
Berichterstatter:	Prof. Dr. William J. Barry Prof. Dr. Bernd Möbius
Tag der letzten Prüfungsleistung:	03.02.2014

Short summary

In this thesis, we employ the state-of-the-art articulatory synthesizer *VocalTractLab* (Birkholz, 2006) for phonetic research and aim to contribute knowledge about fine articulatory details in a range of paralinguistic phenomena of speech. The synthesis experiments cover all major anatomic areas of speech production by investigating topics such as larynx height and associated voice qualities, smiled vowels, vocal age, simulating a laugh and a speech-laugh, and simulating vowels with a Saxon accent which are then integrated into accented words. We also aim to evaluate the synthesizer as a ‘discovery tool’ in phonetic research by checking its versatility, adequacy and quality of output during these experiments in which small articulatory details need to be properly simulated.

From a phonetic point of view, we were able to generate schemata describing the fine articulatory detail of each of the tackled paralinguistic phenomena. Further refinements are subject for future work. Technically we found that, all in all, the basic requirements for articulatory research are met by VocalTractLab, and it proved to be a valuable and flexible tool for basic phonetic research enabling us to design systematic series of experiments regarding the relationship between articulation and acoustics. Its high flexibility can be especially handy when creating expressive speech or accompanying sounds such as laughs, speech-laughs and breathing noises.

Kurzzusammenfassung

In dieser Arbeit setzen wir moderne artikulatorische Sprachsynthese (VocalTractLab, Birkholz, 2006) in der phonetischen Forschung ein, um Fragestellungen zu feinen artikulatorischen Details bei paralinguistischen Phänomenen in der gesprochenen Sprache zu untersuchen. Die Synthese-Experimente decken alle wichtigen anatomischen Abschnitte des Sprechapparates ab, indem wir Themen behandeln wie Larynxhöhe und Stimmqualität, gelächelte Vokale, Alter in der Stimme, Lachen, Sprechlachen, sowie Simulation von sächsischen Vokalen, die anschließend in sächsisch-akzentuierte Wörter integriert werden. Ein zweites Ziel der Arbeit besteht darin zu evaluieren, inwieweit das Synthesesystem als ‚Entdeckungswerkzeug‘ für die phonetische Forschung geeignet ist. Dafür wird seine Vielseitigkeit, Adäquatheit und die Qualität seines Outputs bezüglich feiner artikulatorischer Details untersucht.

Phonetisch gesehen haben wir für jedes paralinguistische Phänomen ein artikulatorisches Ablaufschema entwickeln können, das auf feine artikulatorische Details eingeht. Synthesetechnisch betrachtet konnten wir feststellen, dass die Basisvoraussetzungen für eine Artikulationsforschung mit VocalTractLab erfüllt sind und es sich als nützliches und flexibles ‚Entdeckungswerkzeug‘ erweist, da es erlaubt, systematische Experimente bzgl. Artikulation und Akustik durchzuführen. Seine große Flexibilität ist besonders hilfreich, wenn es darum geht, expressive Sprache oder deren Begleitäußerungen wie Lachen oder Atmen zu simulieren.

Zusammenfassung

In dieser Arbeit benutzen wir artikulatorische Sprachsynthese für phonetische Grundlagenforschung und setzen ein modernes Synthesesystem ein, um artikulatorisch-phonetische Details von paralinguistischen Phänomenen der gesprochenen Sprache zu analysieren. Der Schwerpunkt der Untersuchungen liegt dabei auf einer relativ detaillierten Ebene der Artikulation, die wir die Ebene der ‚feinen artikulatorischen Details‘ nennen. Dieser Begriff soll hervorheben, dass die gängigen phonetischen Beschreibungen von Sprachlauten oft eine Informationslücke hinterlassen. Sie benennen hauptsächlich ausgewählte Aspekte der Lautproduktion oder beschreiben sie auf einem qualitativen und nicht auf einem quantitativen Niveau.

Diese Informationslücke ist besonders prägnant, wenn man paralinguistische Eigenschaften der Sprache beschreiben möchte, da diese normalerweise nicht von gängigen Schablonen erfasst werden, wie sie z. B. in klassischen Phonembeschreibungen verwendet werden. So kann das Phonem /p/ als ‚stimmloser bilabialer Plosiv‘ hinreichend beschrieben werden, der Vokal /aː/ als langer, offener Mittelzungenvokal. Aber wie sieht es aus, wenn man einen gelächelten Vokal beschreiben möchte, und zwar derart detailliert, dass man ihn nachahmen kann? Es gibt zwar Diakritika für gespreizte Lippen bzw. weniger Lippenrundung [a̜ː], doch auf welche Art und Weise werden die Lippen genau gespreizt? Ist diese Spreizung bei allen Vokalen gleich? Welche artikulatorischen Faktoren gilt es außerdem noch zu beachten? Das Beispiel des gelächelten Vokals illustriert nur eines von den Phänomenen, die wir im Rahmen dieser Arbeit behandeln werden.

Dazu führen wir eine Reihe von Synthese-Experimenten durch, in denen maßgeschneiderte Sprache simuliert und anschließend evaluiert wird. Die grundlegende Motivation dazu wird in Kapitel 1 dargelegt. In den Experimenten verfolgen wir zwei Ziele, die miteinander verwoben sind. Zunächst möchten wir dazu beitragen, das Wissen über bestimmte artikulatorische Details bei paralinguistischen Phänomenen zu erweitern. Außerdem möchten wir anhand dieser Experimente ein modernes artikulatorisches Synthesesystem in seiner Funktion als ‚Entdeckungswerkzeug‘ in phonetischer Forschung evaluieren, indem wir seine Vielseitigkeit, Adäquatheit und allgemeine Ausgabequalität bzgl. feiner artikulatorischer Details analysieren. Dieser Versuchsaufbau basiert auf einer Idee der Gegenseitigkeit, d. h. das Werkzeug und die Forschung hängen voneinander ab, beeinflussen einander und jedes

kann eingesetzt werden, um das andere zu evaluieren. Dahinter steht die Annahme, dass ein Synthesesystem grundsätzlich dazu dient, vorhandenes Wissen über Artikulation zu bündeln und zu organisieren, und dass das System uns dadurch ermöglicht einzuschätzen, wie umfassend und adäquat dieses vorhandene Wissen ist. Gleichzeitig kann uns die artikulatorische und akustische Leistung des Systems Anhaltspunkte darüber geben, wie gut seine internen Modelle funktionieren.

In dieser Arbeit setzen wir das Synthesesystem *VocalTractLab* (VTL) von Peter Birkholz (2006) ein. Wir gehen davon aus, dass es, als ein Vertreter der aktuellen artikulatorischen Synthesesysteme, ein durchdachtes technisches Rahmenwerk darstellt, in dem Ergebnisse relevanter Grundlagenforschung auf sorgfältige Weise integriert und implementiert wurden. Da es qualitativ hochwertige Sprache produziert, wie im Demo-Material zu der Doktorarbeit von Peter Birkholz (2006) illustriert wurde, und zudem diese Ausgabe durch das Anpassen artikulatorisch fundierter und transparenter Parameter geschieht, gehen wir davon aus, dass VTL die grundlegenden Voraussetzungen für ein allgemein funktionierendes artikulatorisches Synthesesystem erfüllt. Wie gut es auf der Ebene der feinen artikulatorischen und technischen Details funktioniert, wird Gegenstand dieser Arbeit sein.

Die phonetischen Forschungsfragen der Experimente sind so ausgewählt, dass sie alle Hauptbereiche des Sprechapparates abdecken: Von der Atmung über die Glottis bis hin zum supraglottalen System mit Pharynx, Mundhöhle und Nasenraum. Dadurch werden wir eine variationsreiche Auswahl an paralinguistischen Themen bearbeiten. Dies wiederum fordert alle Hauptbestandteile des Synthesesystems heraus. Bei diesem Vorgehen liegt das primäre Untersuchungsziel im Bereich der Phonetik und die Sprachproduktionsaspekte bestimmen Themenauswahl und Design der Experimente. An zweiter Stelle steht die technische Evaluation des Synthesesystems, die zu einer Bewertung führt, inwieweit es für phonetische Forschung geeignet ist.

In Kapitel 2 beschreiben wir Methoden, mit denen artikulatorische Daten gesammelt werden können. Hierbei stellen wir auch die Sichtweise vor, dass artikulatorische Sprachsynthese benutzt werden kann, um artikulatorische Daten selbst zu generieren. Wir geben außerdem einen knappen Überblick über weitere Arten von Sprachsynthese und wie eng sie sich an dem eigentlichen Sprachproduktionsprozess orientieren. In Kapitel 3 geben wir einen Überblick über VocalTractLab, stellen seine Hauptbestandteile vor und beschreiben, wie es in den Synthese-Experimenten eingesetzt wird. Dabei ist einer der wesentlichsten Bestandteile die gestische Partitur, in der artikulatorische Kommandos auf der Zeitachse definiert werden. In Kapitel 4 bereiten wir den empirischen Teil der Arbeit vor, indem wir die grundlegenden Annahmen, Ziele und Methoden vorstellen, die allen Experimenten gemein sind.

Den Kern der Arbeit bildet dann eine Reihe von sieben artikulatorisch-akustischen Simulations-Experimenten, die mit VTL durchgeführt werden, um jeweils ein phonetisches und ein technisches Ziel zu verfolgen (Kapitel 5 bis 11). Die behandelten Themen sind Larynxhöhe und damit verbundene Stimmqualität, gelächelte Vokale, Alter in der Stimme, Lachen, Sprechlachen sowie Simulation von sächsischen Vokalen, die anschließend in sächsisch-akzentuierte Wörter integriert werden. Zu jedem Thema erarbeiten wir ein mögliches, detailliertes Artikulationsschema, welches von VTL verarbeitet wird und dessen Output anschließend mit akustischen, artikulatorischen und perzeptuellen Mitteln bewertet wird. Gleichzeitig werden verschiedene Module des Synthesesystems getestet, indem analysiert wird, wie realitätsnah die simulierte Anatomie ist und wie gut die Syntheseprozedur funktioniert. Im Folgenden fassen wir jedes der Experimente kurz zusammen.

In Experiment I wird Larynxhöhe in Verbindung mit Stimmqualität variiert. Dabei können wir prinzipielle Effekte, die Larynxhöhe auf Stimmqualität hat, synthesetechnisch nachvollziehen, insbesondere in Zusammenhang mit der Behauchung der Stimme. Wir permutieren alle Kombinationen von relevanten Parametern und erhalten nur dann akustische Messwerte, die der menschlichen Stimmcharakteristik ähnlich sind, wenn Parameterkombinationen vorliegen, die der menschlichen Artikulationsweise entsprechen. In unserem Versuchsaufbau bedeutet dies, dass eine Stimmkonfiguration mit niedriger Kehlkopfhöhe lockerer klingt and von mehr Behauchung begleitet wird als Konfigurationen in neutraler oder gehobener Kehlkopfposition.

Synthesetechnisch gesehen verdeutlicht dieses Experiment, dass VocalTractLab technisch in der Lage ist, subtile und kleinräumige Änderungen im Vokaltrakt und in der Anregungsart in passende akustische Ausgaben umzusetzen. Der Vokaltrakt wird dabei hauptsächlich in seiner Länge variiert, indem der Kehlkopf bzw. ein damit verzahnter Parameter (HY) auf und ab bewegt wird und so erwartungsgemäße Formantänderungen erzeugt werden. Auf ähnliche Weise bewirkt eine Manipulation von Parametern für die Behauchung in der Stimme erwartungsgemäße akustische Änderungen. Die Manipulation von unabhängigen Untersystemen in VTL kann also eingesetzt werden, um Hypothesen zur Sprachproduktion zu testen. Allerdings muss man explizit darauf achten, dass man keine artikulatorischen Konfigurationen produziert, die sich stark von den menschlichen unterscheiden, da das Vokaltraktmodell zur Zeit keinen Mechanismus parat hält, der typische Abhängigkeiten und Ko-Bewegungen zwischen den einzelnen Artikulatoren sicherstellt.

In Experiment II variieren wir vier Vokale so, dass sie als gelächelt wahrgenommen werden, indem wir die phonetischen Parameter Lippenspreizung, Larynxhöhe sowie Grundfrequenz anpassen. Dabei steuert jeder dieser Parameter zum perzeptuellen ‚Lächel-Effekt' bei, allerdings unterscheiden sich die Wirkungen je nach Vokal. Dies beruht z. B. auf der artikulatorischen Grundeinstellung für jeden Vokal, wobei insbesondere das [iː] durch Lippenspreizung nicht als signifikant gelächelter perzipiert wird, da es von Natur aus schon recht gespreizte Lippen enthält. Für den gerundeten Vokal [uː] hat eine Lippenspreizung ungünstigerweise sogar zur Folge, dass anscheinend die perzipierte Phonemqualität verlorengeht. Dies ist ein Anhaltspunkt dafür, dass wir ein noch feinkörnigeres artikulatorisches Schema benötigen, um perzipiertes Lächeln in Vokalen verlässlich induzieren zu können. Ein solches Schema hängt allerdings von den technischen Möglichkeiten ab, die das Synthesesystem bietet.

Synthesetechnisch gesehen variieren wir horizontales Lippenspreizen (LP), Larynxhöhe (HY) und Grundfrequenz (f_0), um die gelächelten Vokale herzustellen. In akustischen Analysen erkennt man, dass alle drei Parameter die theoretisch erwartbaren akustischen Konsequenzen induzieren. Allerdings wird für das [uː] deutlich, dass ein einfaches Lippenspreizen mittels des Parameters LP die phonemische Qualität verletzt, so dass der Laut nicht mehr als ein Vertreter von /uː/ erkannt wird. Daher scheint es sinnvoll, einen zusätzlichen Parameter in VTL zu integrieren, welcher z. B. ein Zusammenpressen der Lippen in der Nähe der Mundwinkel simuliert, während die Lippen gespreizt bzw. zurückgezogen werden.

In Experiment III erarbeiten wir ein komplexes Arrangement an Stimmqualitätseinstellungen, um Stimmen aus drei verschiedenen Altersgruppen zu simulieren (‚jung‘, ‚erwachsen‘, ‚alt‘). Zu den manipulierten phonetischen Parametern gehören Grundfrequenz, eine Behauchungskomponente und eine Komponente für Jitter und Shimmer (‚Rauigkeit‘). Die ‚alten‘ Stimmen werden verlässlich als solche erkannt, wahrscheinlich wegen der deutlich rauen Stimmeigenschaften. Die übrigen Stimmen werden überwiegend als ‚erwachsen‘ eingestuft. Dies ist ein Indiz dafür, dass das vorgeschlagene Simulationsschema ‚alte‘ Stimmen erfolgreich beschreibt, während es ‚erwachsene‘ vs. ‚junge‘ Stimmen noch nicht deutlich genug unterscheidet. Das stärkste perzeptuelle Signal für die Altersgruppenzuordnung scheint in unserem Schema derzeit die Grundfrequenz zu sein.

Synthesetechnisch betrachtet manipulieren wir die artikulatorischen Parameter Stimmlippenabstand, Einsatz von glottalen Lecks, vertikale Phasendifferenz der Stimmlippen sowie Larynxhöhe. Neben diesen artikulatorischen Merkmalen variieren wir außerdem unmittelbar signalbezogene Parameter, indem wir spezielle Gleichungen für Jitter und Shimmer definieren. Man kann also sagen, dass das artikulatorische Synthesesystem auf eine ‚hybride‘ Art und Weise benutzt wird: Die Manipulation der üblichen artikulatorischen Merkmale führt über die Artikulation zu einer entsprechenden Akustik, während eine Manipulation von f_0 und Atemdruck auf unmittelbare Weise auf die Akustik wirkt, ohne dass man die eigentlich involvierten artikulatorischen Akteure spezifiziert, deren Veränderungen dann mittelbar die Akustik beeinflussen würden. Diese ‚hybride‘ Nutzung von VTL wird dadurch ermöglicht, dass man die Grundfrequenz und den Atemdruck unmittelbar durch diesbezügliche physikalische Werte (in Hz bzw. kPa) kontrollieren kann. Insgesamt sind die umfassenden altersbezogenen Manipulationen nur möglich, weil wir ein spezielles Batch Tool verwenden können, das den Synthetisator über eine Programmierschnittstelle kontrolliert. Dieses Batch Tool erlaubt es, systematisch große Mengen an explorativen Parametereinstellungen zu synthetisieren und zu evaluieren. Nur so können wir für die genannten stimmalterbezogenen Parameter praktikable Wertebereiche ermitteln und diese näher evaluieren.

In Experiment IV imitieren wir einen ‚sing-sang-artigen‘ Lacher (‚song-like‘ laugh) in seiner gesamten komplexen Struktur, also auch etwaige Atemgeräusche, die am Anfang und am Ende auftreten. Die Lacher-Imitation wird von Hörern als ein scheinbar natürlicher Lacher in einem Gespräch akzeptiert. Wenn der Lacher isoliert vorgespielt wird, zeigt sich anhand der Perzeptionsergebnisse, dass eine hohe interne Variabilität in den Lachsilben die perzipierte Natürlichkeit erhöht. Die interne Variation wird durch ein Anpassen der phonetischen Parameter Dauer, Intensität und Grundfrequenz erreicht. Insgesamt scheint das vorgeschlagene Artikulationsablaufschema, das sich stark auf glottale und subglottale Aktivität konzentriert, als eine erste Strategie angemessen, um einen kompletten Lacher inklusive Atemgeräusche von Grund auf zu simulieren.

Synthesetechnisch betrachtet zeigt dieses Experiment, dass man mit VTL die nötigen Artikulationsabläufe simulieren kann, die man für einen typischen Lacher braucht. Dies ist besonders bemerkenswert, da VTL ursprünglich für ‚reguläre' Sprache konstruiert wurde, die sich stark an kanonische Abläufe für einzelne Segmente anlehnt. Trotzdem können auch komplexe glottale und subglottale Gesten erfolgreich eingesetzt werden, um die typische, rhythmische Lach-Struktur zu erzeugen. Allerdings stoßen wir mit diesen Manipulationen auch an Grenzen des Synthesesystems, insbesondere im Hinblick auf den maximalen Atemdruck und die Bewegungsrichtung des Luftstroms. Es wäre daher wünschenswert, höhere Atemdruckwerte benutzen zu können sowie nicht nur egressive, sondern auch ingressive Sprache erzeugen zu können. Abschließend wird in diesem Experiment auch deutlich, dass die gesturale Anordnung auf der Partitur kontextempfindlicher ist als gedacht, besonders wenn man versucht, komplexe Befehlsanordnungen zu koordinieren.

In Experiment V imitieren wir einen kurzen Sprech-Lacher und untersuchen die perzeptiven Effekte von syllabischer Reduplikation und gelächelter Vokalqualität auf den Grad der wahrgenommenen Heiterkeit. Während die syllabische Reduplizierung tatsächlich einen (leichten) Einfluss auszuüben scheint, können wir für die gelächelte Vokalqualität nicht die gewünschte Wirkung feststellen. Allerdings ist die durchgeführte Vokalmanipulation in unseren Stimuli sehr subtil, so dass wir davon ausgehen, dass man gelächelte Vokalqualität dennoch als Teil des Sprech-Lach-Systems betrachten sollte.

Technisch gesehen benutzen wir eine Kombination der Parameter aus Experiment II (gelächelte Vokale) und Experiment IV (die pulsierende Struktur von Lachsilben), erweitert durch die Anforderungen der regulären segmentalen Artikulation beim Sprechen. Diese vielschichtige Anwendung von VTL ist in sehr transparenter Weise machbar, hauptsächlich weil die Parameter artikulatorisch definiert sind und sich intuitiv und flexibel einstellen lassen, um die gewünschten Lach- und Sprach-Komponenten des Sprech-Lachers zu erstellen. Auch in diesem Experiment zeigt sich wieder, dass die gestische Alignierung sensibler und komplexer ist als vorher vermutet.

In Experiment VI erstellen wir sechs lange geschlossene und halb-geschlossene Vokale, die einen sächsisch-akzentuierten Klang haben. Wir adaptieren hierbei die artikulatorische Konfiguration von vorhandenen VTL-Vokalen so, dass sie in der akustischen Ausgabe sehr nahe an die Akustik von aufgenommenen Vokalen herankommen. Anschließend prüfen wir visuell, ob die artikulatorischen Vorschläge plausibel sind, und analysieren dann vokalpaarweise, inwiefern sich das sächsische Artikulationsmuster vom unmarkierten, standarddeutschen unterscheidet. Insgesamt erkennt man in unserer Vokalmenge eine Tendenz zur Vorverlagerung und leichten Absenkung des Zungenkörpers.

Technisch gesehen basiert dieses Experiment hauptsächlich auf dem Einsatz einer VTL-internen sogenannten Formanten-Optimierungsfunktion. Diese besteht im Grunde aus einem restriktiven Akustik-zu-Artikulation-Inversions-Algorithmus. Er übertrifft die Präzision bei weitem, die man durch manuelle supraglottale Einstellungen erreichen kann, wenn man bestimmte artikulatorische Verschiebungen nachbilden möchte. Kombiniert mit einer visuellen Plausibilitätsprüfung scheint diese Funktion ein verlässliches Werkzeug zu sein, um artikulatorische Daten für den ganzen Vokaltrakt auf schnelle und umfängliche Weise zu erstellen.

In Experiment VII bauen wir die sächsischen und standarddeutschen Vokale aus dem vorherigen Experiment in zweisilbige Trägerwörter ein und führen eine perzeptuelle Evaluation durch. Deren Hauptergebnis ist, dass die Vokale insgesamt erfolgreich ihren intendierten Varietäten zugeordnet werden. Das bekräftigt die Ergebnisse bzgl. der artikulatorischen Verschiebungen aus Experiment VI. Desweiteren illustriert dieses Experiment, inwieweit ‚feine artikulatorische Details' in regulärer Wortaussprache unterschiedliche Wahrnehmungen in Hörern auslösen können.

Aus technischer Sicht wird in diesem Experiment sehr viel mit Wortsynthese gearbeitet, trotz fehlender Text-to-Speech (TTS) Funktion in VTL. Es wird eine größere Anzahl an zweisilbigen Wörtern generiert, deren hauptbetonter Vokal systematisch manipuliert wird. Hiermit wird verdeutlicht, dass man VTL trotz manueller Konfiguration der Partituren nutzen kann, um ganze Wörter nach streng definierten artikulatorischen Merkmalen zu generieren. Allerdings wird hierbei auch deutlich, dass man einige Herausforderungen meistern muss, bevor man eine gute Aussprachequalität der Wörter erreicht. Unerwarteterweise weist VTL einen relativ niedrigen Grad an ‚Robustheit' gegenüber koartikulatorischen Effekten auf, so dass einige Segmente zunächst kaum verständlich klingen. Dies können wir optimieren, indem wir Feineinstellungen in den Gesten, im Artikulationsaufwand sowie in den Dominanzwerten der Phone überarbeiten. Langfristig sind vermutlich zusätzliche Änderungen in den akustischen Grundeinstellungen von VTL sinnvoll, insbesondere um die Präzision von Plosivverschlusslösungen zu verbessern. Aktuell bleibt offen, ob man segmentale Verständlichkeit eher durch *a)* optimiertes gestisches Timing und andere Attribute von gesturalen Befehlen erreichen kann, oder *b)* ob primär die technische Beschaffenheit der Simulation und der Synthesemodelle angepasst werden sollte. Vermutlich ist es eine Mischung aus beidem.

Obwohl diese Arbeit ihren Fokus auf paralinguistische Aspekte gesprochener Sprache setzt, bringen wir dennoch eine Begutachtung der Synthese von regulären Wörtern ein (Experiment VII), da eine funktionierende Basisäußerung die Voraussetzung ist, um daran paralinguistische Veränderungen vornehmen zu können. Von daher ist es sinnvoll einschätzen zu können, wie diese unterschiedlichen Eigenschaften in VTL umgesetzt werden können. Die empirische Arbeit mit VTL macht deutlich, wie eng die Ebenen interagieren und dass man normalerweise *nicht* einfach die paralinguistischen Merkmale auf den segmentalen Inhalt aufsetzen kann, weil z. B. ein Ändern der Vokalfärbung für einen sächsischen Akzent dazu führt, dass in den Lauttransitionen wie auch in ganzen benachbarten Segmenten ungewollte akustische Artefakte (Störgeräusche) auftreten.

Im Anschluss an die Darstellung der einzelnen Experimente fassen wir experimentübergreifend die Stärken und Schwächen des verwendeten VocalTractLabs zusammen, wie sie in den artikulatorisch-phonetischen Studien zutagegetreten sind (Kapitel 12).

Insgesamt wurde deutlich, dass die Grundanforderungen für artikulatorische Forschung mit VocalTractLab prinzipiell erfüllt sind. VTL stellte sich als hilfreiches Werkzeug für die phonetische Grundlagenforschung heraus, und das, obwohl es zur Zeit noch nicht an eine TTS-Funktionalität angeschlossen ist. Seine Stärken liegen darin, dass es die Simulation des gesamten Sprachapparats innerhalb eines einzigen Koordinatensystems bereitstellt, dass es alle Artikulationsbewegungen und weiteren Einstellungen, die einer gewünschten Audioausgabe unterliegen, transparent macht und dass es einem den direkten Zugriff auf jeden Parameter einzeln erlaubt. Die Vollständigkeit des Systems zusammen mit der transparenten Natur seiner Parameter ermöglichen es uns, systematische Experimente zum Verhältnis von Artikulation und Akustik durchzuführen.

Im Vergleich zu anderen Synthesearten mag artikulatorische Synthese vielleicht nicht die übliche inhärente Natürlichkeit in den Stimmen aufweisen. Dafür bietet sie aber ein hohes Maß an Flexibilität für die Generierung einer Vielzahl verschiedener Äußerungstypen, ohne dass nachträgliche Signalverarbeitungsschritte notwendig sind. Diese Stärke kann besonders gut zum Einsatz kommen, wenn es darum geht, expressive Sprache oder sie begleitende Lautäußerungen wie Lachen und Atmen zu generieren.

Durch die Bereitstellung des Formanten-Optimierungsalgorithmus, der in Experiment VI für die sächsischen Vokale genutzt wurde, bietet VTL ein sinnvolles Werkzeug, um paralinguistisch markierte Phone zu erstellen, die einem Basisphon in der Standardphonliste ähnlich sind. Der große Nutzen liegt darin, dass wir dadurch Laute erstellen können, die ganz spezifischen akustischen Vorgaben, z. B. in der Formantstruktur, gerecht werden. Eine gezielte Erstellung neuer Laute auf manueller Basis war uns nicht möglich, da die manuellen Anpassungen der Artikulatoren zu ungenau und grobkörnig waren.

Bezüglich der Schwächen der benutzten artikulatorischen Synthese lässt sich Folgendes zusammenfassen: Um artikulatorische Adäquatheit und eine hohe Qualität der Sprachausgabe sicherzustellen, sollten einige technische Aspekte bewusst beachtet werden. Man sollte explizit darauf achten, dass man nur natürlicherweise mögliche artikulatorische Einstellungen auswählt, da VTL selbst zur Zeit keinen Mechanismus enthält, um unnatürliche Vokaltraktformen zu verhindern, z. B. indem Abhängigkeiten zwischen den Artikulatoren modelliert werden.

Weiterhin sollte man darauf achten, erwünschte feine artikulatorische Details präzise auf der Partitur zu verankern, da das Level an artikulatorischer ‚Robustheit‘ bei Wortsynthese zur Zeit relativ niedrig ist. Anders ausgedrückt: In VTL wird die ganze Komplexität der Artikulation deutlich. Dies betrifft beispielsweise die gegenseitige Abhängigkeit vom artikulatorischen Aufwand eines ersten Segmentes und der Dauer, Lautqualität und perzeptuellen Prominenz eines benachbarten, zweiten Segments. In ähnlicher Weise muss das stark kontext-abhängige Verhalten von einigen Konsonanten beachtet werden, indem ggf. einzelne Phone für spezielle Kontexte umdefiniert werden, so z. B. bei Wortpaaren wie /ˈzuːmən/ <zoomen>, das einen einwandfreien Klang aufweist, vs. /ˈziːmən/ (ohne Bedeutung), bei dem im Frikativ ein Lispeln auftaucht. Hierzu ist allerdings anzumerken, dass VTL eine Reihe von kontext-sensitiven Phon-Definitionen bereithält, welche unseres Wissens aber eher während der Softwareentwicklung als Hilfsphone genutzt wurden. Eventuell könnte man diese jedoch auch gezielt auf gestischen Partituren einsetzen.

Die stellenweise hohe Sensibilität führt zu der größten Herausforderung in VTL, nämlich dass zur Zeit sehr viel Aufwand nötig ist, um gestische Partituren zu erstellen, die mehr als ein paar wenige Lautsegmente umfassen. Dies ist auch ein Grund, weshalb innerhalb dieser Arbeit die Komplexität der erstellten Stimuli nicht über die Wort- oder Kurze-Phrasen-Ebene hinausgeht. Eine zusätzliche Herausforderung besteht darin, dass die Qualität von synthetisierten Wörtern recht stark von den enthaltenen Lautsegmenten abhängt, was zu einer Einschränkung in der Wortwahl führt. Eventuell würde ein höheres Maß an koartikulatorischer ‚Robustheit' helfen, um schneller Partituren zu erstellen, da man sich kleine ‚Ungenauigkeiten' erlauben darf. Zur Zeit lässt sich der Prozess der Partiturerstellung v. a. durch den Einsatz eines regelbasierten „Song"-Partitur-Formates beschleunigen.

Des Weiteren wurden in den Experimenten ein paar Grenzen bzgl. der zur Verfügung stehenden Parameter deutlich. Dies betrifft z. B. den Wertebereich des Atemdrucks für Lacher und die Kontrolle der Lippenform bei gelächelter Sprache. Größere Wertebereiche und ggf. zusätzliche Parameter könnten hier sinnvoll sein. Abschließend wäre es wünschenswert, ingressive Sprache generieren zu können.

Als Ergänzung zu den Experimenten, die in dieser Arbeit präsentiert wurden, fassen wir in Kapitel 12 eine weitere mit VTL durchgeführte Experimentreihe kurz zusammen. Sie soll illustrieren, dass man mit VTL nicht nur primär artikulatorische Fragestellungen bearbeiten kann. Mit dem Wissen und der Erfahrung aus den oben zusammengefassten Experimenten können wir kurze Phrasen-Stimuli herstellen, die erfolgreich in ‚regulären' linguistischen Experimenten zum Einsatz kommen. Hierbei geht es um die Erforschung perzipierter Unsicherheit in fiktiven Mensch-Maschine-Dialogen, die in 13 Gruppenhörtests insgesamt 390 Teilnehmenden präsentiert werden. Die Stimuli werden dabei mit VTL maßgeschneidert erstellt, wobei auch etwaige extreme Anforderungen wie eine sehr hohe Grundfrequenz umgesetzt werden können, ohne Syntheseartefakte zu erhalten, die typisch für andere Synthesearten sind.

Der Rest von Kapitel 12 befasst sich mit Neuerungen, die zwischenzeitlich in VTL durch Peter Birkholz implementiert wurden, und stellt in knapper Form dar, wie diese zu unseren empirischen Ergebnissen in Relation stehen. Sie umfassen Stimmprofile und Stimmlippenmodelle, Änderungen bei den Vokaltraktparametern, Einführung eines neuen Dominanz- bzw. Interpolationsmodells, sowie die Simulation von bestimmten Segmenten. Insgesamt haben die Neuerungen einen positiven Effekt auf den Arbeitsprozess und die Einfachheit, mit der man gutklingende Äußerungen erstellen kann, ohne dass sie die empirischen Ergebnisse unserer Artikulationsexperimente signifikant verändern würden, wenn man sie mit VTL 2.1

wiederholen würde. Einige der Änderungen, die jetzt systematisch in VTL integriert sind, haben wir bereits auf ähnliche Weise in Form von punktuellen Hilfskonstruktionen in unseren Experimenten eingesetzt, wie z. B. die kontextabhängige Dominanz-Modellierung in Phonen. Eine charakteristische Stärke von VocalTractLab, nämlich die Möglichkeit, jeden artikulatorischen Parameter einzeln zu manipulieren, ist in VTL 2.1 genauso vorhanden wie in VTL.

Kapitel 13 schließt diese Arbeit mit einem Fazit und Ausblick ab. Die vorliegende empirische Arbeit hat unser Wissen über feine artikulatorische Details in einer Reihe von paralinguistischen Phänomenen erweitert, die mit VocalTractLab simuliert wurden. Die Experimente haben außerdem den derzeitigen Stand von moderner artikulatorischer Synthese illustriert und ihre Einsetzbarkeit in der artikulatorisch-phonetischen Forschung aufgezeigt.

Acknowledgments

The years during which I worked on this thesis have been a time of personal and professional growth. First of all, I would like to thank Prof. emerit. Dr. William J. Barry for the inspiring supervision, yielding many fruitful discussions and insights, and for making me tackle the topic of Artikulationsbasis in regional accents, which proved to be a very inspiring field of work to me. I would also like to thank Prof. Dr. Bernd Möbius for valuable feedback, and advice for the writing process.

Through a series of conversations with Prof. emerit. Dr.-Ing. Wolfgang Hess and Prof. Dr. Bernd J. Kröger, I got introduced to Dr.-Ing. Peter Birkholz and his work on the articulatory synthesizer VocalTractLab. This resulted in several inspiring meetings and continuous collaboration and support over the years for which I am deeply grateful. Special thanks also go to Dr. Jürgen Trouvain, who encouraged me to combine issues on laughter and smiling with articulatory synthesis. I would also like to thank Prof. Dr.-Ing. Elmar Nöth for introducing me to the topic of vocal aging.

Special thanks also go to the staff members and fellow Ph.D. students of the International Research Training Group *Language Technology and Cognitive Systems*. Last but not least, I would like to thank my colleagues, family and friends who supported me over all the years.

Contents

Chapter 1

Introduction

"A smile often says more than a thousand words."

1.1 Motivation

Human speech is very efficient because it not only transmits linguistic content, the meaning of the words, but also information about the speaker and the situation. This is the case even when this information is meant to be irrelevant, as in national news broadcasts. The speech of even the most neutral of the professional news readers reveals something of their individuality and the role they are playing in that particular situation. Listeners are used to this kind of extra information and expect it, too. If the personal and situational information is missing, a voice tends to sound monotonous or boring.

Historically, there has long been a wish to understand the phonetics behind the mechanisms of the human voice which generate the different levels of information transmitted in an utterance. Besides the intended linguistic content, these include personal aspects, i.e. biological attributes and speaking habits, such as gender, age and geographical origin as well as situationally linked features expressing emotion, attitude or status, such as laughing, smiling or the expression of dominance through a 'big' or 'deep' voice. All these features are considered to be 'paralinguistic' or 'extralinguistic'. The concept of paralanguage[1] (cf. Trager, 1958, 1961; Pittenger et al., 1960) has been addressed under different names and with slightly different definitions. In this thesis, the term 'paralinguistic' is seen in opposition to the *linguistic* aspects of what is said. All aspects relevant to spoken communication that do not belong to the formal linguistic system are regarded as 'paralinguistic' and there-

[1] "phenomena, which accompany language, (...) are now handled under the term paralanguage." (Trager, 1958: 8)

fore this term will be used throughout the thesis. A further distinction between the different phenomena is not the focus of this work.

The paralinguistic phenomena are an important, even essential part of speech communication since they are what makes utterances natural and particular to an individual. However, their mechanisms are not fully understood yet. Nevertheless, in recent years, knowledge about the phenomena has increasingly been applied to try to make computers talk in a comparably rich, pleasant, natural, and appropriate way, opening up the field of expressive speech synthesis. As commercial systems have generally mastered the task of synthesizing intelligible speech, and their areas of application have increased, paralinguistic facets of human voices now need to be modeled as well because we cannot express everything merely through the words we use. It is also important how we pronounce the words and what we add around them. But before we can master these additional demands on synthesis, we need to know more about the details of their articulatory-phonetic foundations.

One type of synthesis which is particularly suited to dealing with the great variability of expressive speech is articulatory speech synthesis. Of course nowadays there are methods of synthesis that sound better than state-of-the-art articulatory synthesizers and are more advanced in terms of automated processing to the point of business applications. However, the constitutive advantage of articulatory synthesis is its flexibility. Furthermore, it is a method of synthesis that has always been very attractive for conducting research into articulatory problems because all its parameters are to a high degree transparent in an articulatory sense. Although it still has a long way to go, articulatory synthesis seems to be the most economical type of synthesis, if we understand the speech production details and their perceptual effects on listeners.

In this thesis, we take up the idea of using articulatory synthesis for phonetic research and employ a current state-of-the-art articulatory synthesizer to investigate phonetic problems by conducting a series of experiments in which synthetic speech is simulated and evaluated. The focus of the speech investigations is located on a relatively detailed level of articulation, which we later introduce properly under the term of ‘fine articulatory detail’. The term is supposed to highlight the fact that phonetic descriptions of speech sounds often leave an information gap because they only specify selective aspects of sound production or describe it on a qualitative and not a quantitative level. This gap is particularly noticeable when describing paralinguistic properties of speech because they are rarely covered by established patterns of articulatory description such as those used for the phonemes of a language. ‘Fine articulatory detail’ will also be used as a complement to ‘linguistic’ aspects of speech, i.e. the lexical meaning of the words, in which the presence of a phoneme as a whole is considered, and not so much the articulatory details of its production.

With the experiments in this thesis, we pursue two inter-related goals. Firstly, we aim to contribute knowledge of fine articulatory detail about a range of paralinguistic properties of speech with the help of synthesis experiments. Secondly, we aim to evaluate a state-of-the art articulatory synthesizer as a ‘discovery tool’ in phonetic research by testing its versatility and quality of output during these experiments in which small articulatory details need to be properly simulated.

In other words, the basic principle underlying the experiments presented in this thesis is ‘mutuality’. On the one hand, we work in the style of a traditional line of research (cf. e.g. Carlson and Granström, 1997) which assumes that speech synthesis can be deployed as a phonetic tool. It is used to study the effects of individual parameters on auditory perception. At the same time, on the other hand, “considerable understanding of the speech act itself” (Carlson, 1995: 9932) is necessary to create articulatory synthesis in the first place. Thus, the tool and the research depend on each other, influence each other, and each can be used to evaluate the other. In a sense, a synthesizer serves as a means to bundle and organize existing knowledge and to challenge the quality of the implemented knowledge, while at the same time the articulatory performance and the speech output of the synthesizer indicate how well its internal models work.

In this thesis, we use the synthesizer *VocalTractLab* (VTL), which has been developed by Peter Birkholz (Birkholz, 2006). We assume that, as a current state-of-the-art articulatory synthesis system, VTL represents a well-grounded technical framework in which the results of relevant basic research in speech production and the resulting models have been integrated and implemented in a careful manner to obtain the complete system. The fact that we can produce high-quality speech output by adjusting articulatorily defined parameters, as can be heard in the demo material provided in Peter Birkholz’s doctoral thesis (Birkholz, 2006), we take as indication that the basic requirements for a functioning articulatory synthesizer are met overall. The status of VTL regarding articulatory and technical *details* will be the subject of the experiments in this thesis.

The phonetic research questions addressed in the experiments are selected in such a way as to cover all major areas of speech production, ranging from pulmonic features via glottal activity to supraglottal action in the pharyngeal, oral, and nasal parts of the vocal tract. This challenges all core modules of the synthesizer. In doing so, we place the primary goal of the thesis in the phonetic domain and phonetic questions determine the topic selection and design of the experiments. The technical evaluation of the synthesizer, i.e. how well it is suited to phonetic research, constitutes the secondary goal of investigation.

In the following we delineate the specific research aims and give an overview of the thesis structure.

1.2 Phonetic research aims

We conduct a series of experiments that focus on the paralinguistic properties of speech communication and point out articulatory details of their production. The phenomena we investigate are larynx height and voice quality, ‘smileyness’ in vowel articulation, traits of age in the voice, occurrences of laughter and laughed speech, and articulatory details of Saxon-accented vowels and their perception when integrated into words.

While general phonetic ideas and descriptions of these phenomena are available, their specific properties in terms of fine articulatory detail are not fully understood yet. The questions that arise are: How can these phenomena be described in a precise way? How can we obtain suitable articulatory data? How can we manipulate them and thus study the role of their individual properties?

To answer these questions, we propose an articulatory scheme of how the paralinguistic information of each phenomenon can be simulated. The initial content of the schemata is based on existing knowledge reported in the literature, adapted to the parametric dimensions of the synthesizer. The schemata use the control features offered by VocalTractLab, notably the gestural score and the storage of articulatory phone configurations, and describe the phenomena in terms of the temporal succession of events and in terms of precisely defined shapes of the articulators. With the execution of these descriptions during simulation we obtain the corresponding articulatory data and by manipulating the commands, we can study the effects of the individual factors on the acoustic, articulatory and perceptual level.

It is important to note that the articulatory suggestions we obtain only present one possible answer to our research questions since every acoustic speech-related event can be generated by an indefinite number of different actions and settings in the vocal tract, in humans as well as in articulatory synthesis. However, some are more plausible or more often used than others.

The aspects of speech production we are interested in are limited to the observable movements of the articulators. This is sometimes called the kinematics of speech production. Anything that is related to brain activity in a neurological sense is not within the scope of this work. We only refer to ‘commands’ in a technical way, denoting the articulatory gestures that are used to produce desired articulatory movements in space, and providing a means of describing the temporal unfolding of articulatory events.

The main result in answer to the phonetic research questions is the development of articulatory schemata that successfully simulate the above-mentioned paralinguistic phenomena of speech production. These findings can help to better understand minute aspects of speech production and the relatively strong effects they have on listeners. Furthermore, the results may support efforts to make speech synthesis, and articulatory synthesis in particular, sound more naturalistic, by expanding the articulatory-phonetic foundation of the speech simulations.

1.3 Technical assessment of VTL

The technical goal of the thesis is to illustrate how VTL can be used as a ‘discovery tool’ for articulatory speech research, how reliable it is, what its strengths are and where challenges reside.

We basically employ VTL as a data acquisition tool, acquiring a comprehensive, multidimensional record of the articulatory activity generating the acoustic signal. In terms of the traditional phonetic experiment, VTL thereby replaces the speakers to record, the articulatory sensors and the digitization of the recorded articulatory and acoustic data (cf. Section 2.5). A general strength of this method is that all the articulatory data are provided in one single coordinate system spanning the whole speech apparatus.

In the technical evaluation we focus, firstly, on the fine-grained modeling of different anatomical areas relevant in speech production and their acoustic effects, and secondly, on the different aspects of the synthesis procedure, especially its overall robustness and the effects of individual synthesis parameters. Other ways of evaluating an articulatory speech synthesizer are also conceivable. While we conduct an evaluation that is focused on the ‘low-level’ articulatory behavior of the system, other potential topics of evaluation could include the linguistic performance (intelligibility, adequateness, pleasantness etc.) e.g. in different speaking styles (such as reading style vs. conversation), or an evaluation of synthesis quality vs. system flexibility vs. computational performance in terms of resources needed (computational capacities, time, footprint). These aspects of synthesizer performance will not be considered in this thesis since they lie outside the domain of the dual goals of phonetic and technical evaluation described above.

We will also not go into fundamental technical details about the synthesizer’s internal models. We would only like to observe that every model needs to make simplifications and is subject to possible technical or other restrictions such as computing power. Therefore, it has to be expected that certain restrictions within the synthesizer and regarding its applicability as a phonetic tool will apply. Furthermore, any automation, such as in the direction of automated word synthesis, is beyond the scope of this thesis. Instead, we aim to provide insight into the possibilities to (manually) control fine articulatory detail to obtain smooth acoustic rendering of specific speech phenomena.

The main result of the technical evaluation in this doctoral research project is that VTL can indeed be considered a sophisticated tool for articulatory speech research. The broad basis of the simulation framework works well. However, there are local limitations in some production areas, e.g. regarding consonant pronunciation and co-dependencies between model articulators, and the amount of manual work needed to build an utterance is relatively high.

1.4 Structure of the thesis

At the core of the thesis, we present a series of seven experiments which are conducted using VocalTractLab to pursue a phonetic and a technical evaluation goal. Before reporting on the experiments, the first three chapters present background information that is relevant to all experiments.

Chapter 2 provides an introduction to research dealing with articulatory data, and introduces the idea of using articulatory synthesis as a tool for articulatory data acquisition. It also gives a brief overview of different types of speech synthesis, with regard to their proximity to the actual speech production process. In Chapter 3, we give an overview of VocalTract-

Lab (VTL), i.e. the articulatory speech synthesis system that we use in all our experiments. Its main components are introduced and we describe how VTL is used in the experiments. Chapter 4 sets the stage for the empirical part of the thesis by describing the assumptions, aims and procedures common to all experiments.

Chapters 5 through 11 then each present an articulatory-acoustic simulation experiment using VTL, dealing with the topics of larynx height and associated voice qualities, smiled vowels, vocal age, simulating a laugh and a speech-laugh, and simulating vowels with a Saxon accent which are then integrated into accented words.

In Chapter 12 we summarize the main empirical findings with regard to phonetic and technical aspects, including a cross-experimental assessment of the tasks and further work done with and on VTL. Chapter 13 completes the thesis with final conclusions and an outlook.

Appendix Sections A to F provide additional details on technical synthesis settings and supplementary information for a number of experiments.

Chapter 2

Obtaining articulatory data and simulating articulation

In this thesis, we investigate fine articulatory detail in paralinguistic speech phenomena. We use the term *fine articulatory detail* with reference to the concept of *fine phonetic detail*, introduced by John Local (see e.g. Local, 2003). Fine articulatory detail shall refer to mostly subtle settings and movements of the articulators. Similar to Smith (2004: 12), the term 'detail' implies that the cues under investigation can only be found by a close analysis of the articulation. This requires a precise description framework. In particular we want to raise awareness that for some speech phenomena millimeters or milliseconds matter. The variation in articulatory detail may result in subtle acoustic differences, but also in very audible differences ("quantal nature" of speech, cf. Stevens, 1989).

Investigations on such a fine-grained articulatory level depend on equally fine-grained articulatory data to work with. In this chapter we provide some background on ways of obtaining articulatory data, and possibilities to use them in simulations or speech synthesis. The chapter is intended to serve as a foundation for the introduction of the articulatory speech synthesizer VocalTractLab (VTL, Chapter 3), and for the methodology used in our experiments (Chapter 4).

We acknowledge the achievements of today's tools by sketching the long way they have come since the late 19th century (Section 2.1). Afterwards, we discuss how the data can be used in articulation simulation and speech synthesis: In Section 2.2, we provide a brief background on articulatory inversion, and in Section 2.3, we present an overview of different speech synthesis methods, focusing on their relation to speech articulation. In Section 2.4, we discuss general characteristics of articulatory synthesis, before pointing to a perspective in which articulatory synthesis is regarded as a data management framework and articulatory data generator (Section 2.5). VocalTractLab represents such a framework, and we actually employ it as a data generator in our experiments.

2.1 Articulatory phonetic research methods

Speech articulation is very complex. It involves many organs of the human body and relies on the accurate coordination of the different physiological components. Aspects of lung volume and pulmonic pressure have to be co-ordinated with glottal and supraglottal settings to produce a source signal, and with supraglottal configurations and movements, which modify the source signal to shape the actual speech sounds. Relevant details of articulation will be discussed in Section 3.1.1 in the next chapter, in conjunction with their representation within the articulatory speech synthesizer VocalTractLab. In the present section, we focus on how the articulatory settings and movements can be captured. We begin with a short retrospective, to illustrate the long road of technological development, and end by sketching the technical capabilities of state-of-the-art devices.

2.1.1 The approach of self-observation

Aside from the influential work of the Sanskrit grammarian Pāṇini from about 2500 years ago (Böhtlingk, 1887), the linguistic and phonetic accounts on speech production have become more and more numerous since the late 19th century. They largely describe speech production as captured by self-observation, careful listening and reproduction. Descriptions of sound productions were generally impressionistic in nature, exemplified by the following example about the “peculiarities” of the “general character of English speech”: “The tongue is broadened and flattened, and drawn back from the teeth (...), and the fore part of it is hollowed out, which gives a dull sound, especially noticeable in *l*.” (Sweet, 1890b: 4)

Speech scientists used their bare hands to analyze speech production details by localizing and carefully touching e.g. the thyroid cartilage and following its movements while producing different sounds (Leky, 1917: 89). The impressionistic descriptions were useful to gain a basic understanding of articulation, and could be applied e.g. to phrase instructions for ‘proper’ pronunciation. Accordingly, the articulatory descriptions were usually meant to support foreign language learners, speech training and elocution classes, public speakers, language researchers (Sweet, 1877, Sievers, 1901), or people with speaking or hearing impairments (Bell, 1867).

While speech articulation can be described on a *qualitative* level using proprioceptive experiences, for research and simulation (as in speech synthesis) it is important to also obtain precise data on a *quantitative* level: A precise description of each part of the vocal tract at any point in time. This is the basis for sophisticated simulations. To obtain these precise descriptions, many different tools have been developed, mostly in the area of medical applications. After having undergone a long history of tool development (Section 2.1.2), state-of-the-art devices can nowadays make explicit how minimal articulatory differences may have a critical impact on a sound.[1] In order to make the grade, the tools have to meet high standards of precision, particularly regarding temporal and spatial resolution (Section 2.1.3).

[1] Combined with or supported by acoustic theories of sound production, such as Fant (1960); Stevens (1989).

2.1.2 Early instrumental techniques

It was during the second half of the 19th century that the empirical methods for analyzing speech physiology became very diverse and started contributing substantial knowledge to the area of speech production. Henry Sweet describes these contributions as coming from "physiologists and physicists rather than practical linguists" (Sweet, 1877: vi), mentioning physiological accounts e.g. by Brücke (1876) and Merkel (1866). According to Wängler (1972: 160), a number of instruments were introduced during that time, such as the labiograph, palatography with an artificial palate, using the manometer to determine the degree of velic closure, and recording X-ray images of movements in the vocal tract.

Although the early instrumental techniques were valuable tools for speech production research, they had three main weaknesses. Firstly, being based on mechanical principles, they did not achieve a very high precision. Secondly, it was generally not possible to record the physiological data or the speech for later analysis. So everything had to be evaluated immediately. Thirdly, the tools were in direct contact with the articulating organs and therefore influenced the process of articulation.

In the 20th century, progress was made due to electronic support within the tools and the development of other signal recording and processing techniques. Traditional mechanisms were optimized, and new ones were invented. Descriptions can be found e.g. in Borden and Harris (1984) or Wängler (1972). Although these techniques represented powerful tools for articulatory data acquisition, their main weakness was that many of them were direct methods, or if indirect, posed health risks due to radiation. Direct methods, in contrast to indirect methods, influence the articulation because they are in direct contact with the articulators, such as pellets or artificial palates. Indirect methods record data remotely (cf. e.g. Stone, 2010). Another problem was that most of them could only track small parts of the vocal apparatus. Finally, the data were usually not stored in digital form, so it was difficult to post-process them or to use them in more comprehensive data representation frameworks.

In Borden and Harris (1984), the authors state some 30 years ago, mainly related to supralaryngeal movements: "Many of the available techniques cannot be used for tracking movement of many points simultaneously. (...) The development of a more adequate technology for speech movement analysis is a widely recognized need." (Borden and Harris, 1984: 249)

2.1.3 State-of-the-art instrumental techniques

The last three decades have indeed brought about a noticeable change in the possibilities to study speech and record articulatory data, and the "physiological measurements have improved at an extraordinary pace (...) revealing inter-articulator relationships that could only in the past be addressed theoretically." (Stone, 2010: 10) One of the main attributes of the state-of-the-art techniques is that they comprise indirect methods which work in the three-dimensional space. Therefore, comprehensive data can be collected remotely with little to no influence on the articulation of the speaker. Due to the digitization of the data, they can be post-processed extensively, if suitable software is available, increasing the amount of physiological data available to study speech production and to model vocal tracts (cf. Stone, 2010: 10). Other key attributes of modern tools include faster sampling rates, higher spatial resolution, less influence on the naturalness of speech production, and minimized health risks.

We give a short overview of different methods of articulatory data acquisition before discussing some challenges associated with them. This account is intended to provide a picture of the data basis that enables state-of-the-art articulatory synthesis systems.

Stone (2010) portrays three groups of devices, namely state-of-the-art imaging techniques, point-tracking devices, and tools for measuring tongue-palate interaction. Among the imaging techniques we basically find X-ray, computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound. In the group of point-tracking devices we find tools that track individual points of the articulators by localizing the pellets that are glued to them, such as EMA (Electromagnetic articulography) and X-ray microbeam. The third group of devices is specialized on tongue-palate interaction, mainly referring to electropalatography (EPG). Other techniques used to record the shape of the vocal tract and movements therein include plaster casts, transillumination, fiberscope filming and optopalatography (Maeda et al., 2008: 7, Ridouane et al., 2006). Table 2.1 provides an overview of methods available for articulatory data acquisition as discussed in Maeda et al. (2008). The table focuses on the nature of the data and limitations regarding the application of the technique. For more information cf. Maeda et al. (2008). A very useful overview concerning flesh-point tracking techniques, ultrasound tongue imaging (UTI), and electropalatography can also be found in Mennen et al. (2010).

There are a number of general challenges that all tools are exposed to. Capturing the movements within the vocal tract is a complex task, with technical challenges arising from both the nature of the speech movements and the dimensions and the shape of the vocal tract. The speech movements are very fast and highly variable. To adequately capture the movements of articulators, one would need “60 frames/s to observe muscular-force induced articulatory movements, while 1000 frames/s would be required to observe aerodynamic-force induced movements, such as those during consonantal release” (Maeda et al., 2008: 14). The high variability of speech represents a special challenge for techniques that need multiple repetitions of one articulation because they may produce ‘averaged’ articulation results which do not adequately mirror the real articulation. Furthermore, in some regions of the vocal tract, very small deviations in shape cause strong acoustic variation while other areas are not so sensitive (quantal nature of speech, Stevens, 1989). This calls for recording techniques with very high spatial resolution. Lastly, most parts of the vocal tract are not visible from the outside, and back parts may be covered by more anterior parts This limits the spatial coverage of many methods.

Table 2.1: Different techniques for articulatory data recording. Adapted and extended from Maeda et al. (2008: Table 2.1 and corresponding text sections).

	Electro-magnetic Articu-lography (EMA)	**Magnetic Resonance Imaging (MRI)**	**Ultrasound**	**X-ray**	**X-ray Microbeam**
Entire vocal tract	No	Yes	No	Yes	No
Temporal resolution	200 Hz	0-24 Hz	30-200 Hz	50 Hz	40-160 Hz
Spatial resolution	Best if close to midsagittal plane	Stationary: Okay, Real-time: Rather low quality (blurry)	Hard to detect edges, and to keep absolute alignment	Good (0.3 mm), but contours hard to detect, multiple layers, shadows	Good, and pellets are easy to trace
3D	No	Yes	No	Yes	No
Affects articulation	Yes (Pellets)	(Yes) (Supine position)	No	No	Yes (Pellets)
Health risks	No	No	No	Yes	Yes
Portable	No	No	Yes	No	No
Low-cost	No	No	Yes	No	No

Captured data need to be post-processed to unfold their full value for research. When processing the data, the main challenge lies in the fact that the amount of collected data is usually huge and may be distorted by noise and errors. Many techniques record low-level, raw image data, which need to be cleaned and transferred into higher-level articulatory or spatial data. This requires extensive data post-processing techniques, so-called vocal tract image processing techniques (Maeda et al., 2008: 23), involving image segmentation, smoothing, and interpolation. Additional challenges may arise from normalization issues across different speakers (Mennen et al., 2010), and also from effects of the posture of the speaker during the recordings (Tiede et al., 2000, Steiner, 2010). This refers to differences in articulation between supine and upright positions due to gravitation (taken up in more detail in Section 10.4.2, p. 163).

Mennen et al. (2010: 35) point out in their review of tools that

> “articulatory techniques have their disadvantages. They can be costly, time consuming, mislead the naive researcher, provide only partial information on particular aspects of articulation and, finally, quantitative analysis from images (ultrasound or X-ray) can be problematic. (...) Moreover, articulatory techniques are more intrusive than simple audio recordings, meaning that their presence may impede natural speech production or make participants harder to recruit. Finally, software and free corpora are only just becoming more widely available.”

As Stone (2010: 31) sums up: “Instrumental studies of physiology are challenging and, no single instrument provides total vocal tract information.” We need the data, however, to further our understanding of speech physiology, speech disorders, and coarticulation strategies, and for testing of theories and models (Stone, 2010: 31). Although not every instrumental method is equally strong in every area, this is not a real caveat because the articulatory data can serve as a valuable basis for the development and evaluation of other methods. Two kinds of them will be discussed in the subsequent sections.

Firstly, in a bird’s eye view, the individual articulatory data sets can be merged within comprehensive frameworks. The accumulated data can thus be made available for analysis and simulation of speech movements, and for simulation of articulation and acoustic output. Thus, regardless of the problems and challenges mentioned above, the collected physiological data of the vocal tract build a valuable basis for synthesis frameworks such as VocalTractLab. Combined with models of movements and acoustic simulation, they provide the basis to enable articulation research using speech synthesis. Since not only articulatory synthesis has been used to study articulatory aspects of speech, we introduce different kinds of speech synthesizers in Section 2.3, with special attention given to their proximity to actual speech articulation.

Secondly, the instruments discussed above can also collect valuable data bases for the evaluation of acoustic-to-articulatory inversion algorithms, i.e. mathematical methods for articulatory data acquisition. We discuss aspects of this field in Section 2.2 because an example of such an algorithm is implemented in VocalTractLab, and we use it in Experiment VI to generate vowel sounds (Chapter 10). In Sections 2.4 and 2.5, we return to articulatory synthesis in more detail.

2.2 Data generation by inversion

This section deals with aspects of acoustic-to-articulatory inversion, also named inversion mapping (Richmond, 2007), acoustic-to-articulatory (A-to-A) mapping, or articulatory inversion (Neiberg et al., 2008). Inversion aims at estimating or recovering the “underlying sequence of articulatory configurations” (Richmond, 2007: 68) which produced a particular acoustic speech signal. It therefore denotes the process of calculating the geometric shape that can produce a given acoustic output. According to Neiberg et al. (2008), it is “one of the fundamental problems in understanding speech production.” It is particularly challenging because “multiple evidence exists to suggest the articulatory-to-acoustic mapping is many-to-one, which means that instantaneous inversion of this mapping results in a one-to-many mapping.” (Richmond, 2007: 68, cf. also Schroeter and Sondhi, 1994.) Strong evidence is e.g. provided by the early bite-block experiments and their further considerations for speech production (cf. e.g. Gay and Turvey, 1979; Perkell, 1979), where perturbed and therefore compensatory articulation still yielded speech within a normal acoustic range. Likewise, early computer simulations illustrate how a given sound is generated using “many different vocal-tract shapes” (Atal et al., 1978: 1535).[2]

A wide range of methods has been applied to calculate inversion mappings. They include mathematical models of speech production, also called analytical approaches (Richmond et al., 2003), articulatory synthesis models, and, more recently, machine learning models applied to recorded articulatory data, including artificial neural networks, codebook mapping methods and Gaussian Mixture Models (as summarized in Richmond, 2007: 68), neural networks for analysis-by-synthesis paradigms (Shirai, 1993), multilayer perceptrons, mixture density networks, or trajectory mixture density networks (Richmond et al., 2003; Richmond, 2009). The machine learning techniques are used to minimize the error between the original recorded vocal tract configuration and the configuration which is estimated based upon the characteristics of the acoustics (Neiberg et al., 2008: 1485).

Of special interest here are the articulatory synthesis models. They can be part of an optimization algorithm which tries to minimize a cost or error function by iteratively adjusting model parameters. The generated acoustic output is compared to the original acoustic signal and aims at closely resembling it (cf. Richmond et al., 2003: 153). In this way, articulatory movements are estimated from the target speech signal, and furthermore, new articulatory movements can be generated, based on a phonemic transcription (Shirai, 1993). Currently, VTL is being integrated in such a training setup using a stochastic gradient method to optimize the articulation to match natural utterances (Prom-on et al., 2013).

For working on articulatory topics, inversion mapping techniques have the advantage of providing articulatory data in a rather “convenient” manner (Richmond, 2007: 67) compared to the instrumental techniques for data collection. Although the collected data are helpful, “they are still invasive techniques and require bulky and expensive experimental setups. Therefore, there is interest in developing a way to recover an articulatory representation from the acoustic speech signal” (Richmond, 2007: 67). Aside from supporting fundamental research on speech articulation, convenient access to, and fast generation of, articulatory data has a range of applications in the field of speech processing. These include low bit-rate speech coding, speech analysis and synthesis, especially articulatory speech synthesis, robust automatic speech recognition, and animating visual agents such as talking heads (Neiberg et al., 2008; Richmond, 2007).

[2] There seems to be some discussion, however, whether the acoustic-to-articulatory mapping is really many-to-one. This would be the case “if more than one articulatory position can produce exactly the same acoustic features” (Neiberg et al., 2008: 1485). The authors argue that “In real continuous speech (...) the possibility of finding two data points with exactly the same acoustic parameters is abysmally poor.” (p. 1485) The view of the many-to-one relation may be due to a “misleading” quantization of the acoustic space. Thus “if two data points within this [single acoustic] quantization range fall sufficiently apart in the articulatory space, then the mapping is said to be non-unique.” (p. 1485)

2.3 Synthesis methods and their relationship to speech articulation

Speech synthesis has the strength to produce utterances that can be closely tailored to the needs of a given empirical design. However, the degree to which the synthesized utterances can be adapted to a given task or research question varies for each synthesis method. Since we investigate fine articulatory detail, we focus on the relation of different synthesis methods to articulation and physiology aspects. Since most of the methods operate at the acoustic level of speech processing, they seem less suited to the direct study of articulation: They offer mostly an indirect perspective on articulation, with the need to infer articulation from acoustics. In the following, we provide a brief comparative overview of different speech synthesis methods to discuss their general suitability for articulatory research. After the overview, the method of articulatory synthesis is discussed in more detail. On these grounds, we will present, in the next chapter, the articulatory speech synthesizer that we used in our experiments.

2.3.1 Mechanical synthesis

Mechanical synthesis is probably the earliest form of speech synthesis. It denotes a kind of synthesizer which is built from real hardware materials. We can generate speech-like sounds by playing the synthesizer as one would play a musical instrument. One of the most widely known representatives seems to be the Speaking Machine of Wolfgang von Kempelen, developed around 1791 (Brackhane and Trouvain, 2011). It has been replicated at various occasions (cf. e.g. Trouvain and Brackhane, 2010), and could be called the "first ever functioning mechanical speech synthesiser" (Trouvain and Brackhane, 2011a: 162).

Another pioneering way of synthesizing speech was put forward by Christian Gottlieb Kratzenstein a few years earlier by developing organ pipes. Each pipe would synthesize one particular vowel (Trouvain and Brackhane, 2011b). However, it is attributed to Von Kempelen to have been the first one to realize *coarticulation* in his machine and to have integrated the basis for the different speech sounds into one single apparatus, roughly resembling the idea of the human vocal apparatus (Trouvain and Brackhane, 2011b: 166).

Although the individual parts of this synthesizer clearly related to parts of the articulatory tract, they were not primarily articulatory in nature (such as the bellows resembling a simplified version of the human lungs, cf. e.g. Table 1 in Trouvain and Brackhane, 2011a). The main goal was to create sounds that would resemble the acoustic characteristics of speech. Thus the development was guided by acoustic goals and only to a certain extent by articulatory imitation goals.

Modern counterparts of the historical mechanical synthesis machines can be found in systems such as flute playing robots and anthropomorphic talking robots that are electronically controlled and generate (speech) sounds with their real existing body mass (cf. e.g. Fukui et al., 2008, 2010).

Due to the non-virtual nature of the mechanical synthesis and their relatively simple structure, a system such as Von Kempelen's can serve as a useful didactic tool to convey basics of articulation to a layman audience (Trouvain and Brackhane, 2011b). Articulation research, however, is not really feasible because the simple structure allows only basic articulatory patterns.

2.3.2 Formant synthesis

As in mechanical synthesis, formant synthesis also uses general knowledge of vocal tract characteristics, and again the main goal is to simulate the appropriate acoustics by using the *acoustic* properties of the oral and nasal cavities. In order to build a functioning formant synthesizer, one important aspect is the appropriate implementation of formant transitions between sounds. Thus, a general understanding of speech sound production and its relation to acoustics and perception is necessary.

The characteristic feature of formant synthesis is that it combines different acoustic sources, i.e. voicing, fundamental frequency, and noise sources. Since the entire output is created by sound models, the speech is highly controllable. It can be flexibly used e.g. to create different intonation patterns, which can serve as the basis for emotionally rich speech.

Although formant synthesis voices usually sound artificial, their strength lies in their high degree of intelligibility and the possibility to increase speech rate for faster access of written content, e.g. when using a screen reader (cf. e.g. Moos and Trouvain, 2007). Knowledge about the vocal tract is used implicitly only, in order to model acoustic properties, and is not used to model any articulatory gestures (Taylor, 2009). Formant synthesis, though, is sometimes combined with (pseudo-)articulatory parameters to undertake articulation research (see Section 2.4).

2.3.3 Concatenative and statistical-parametric synthesis

While mechanical and formant synthesis produce their speech output from scratch, data-driven techniques such as concatenative synthesis use small segments of natural utterances that are pre-recorded. Thus, the voices tend to have a high degree of naturalness because the original voice characteristics are stored in the recorded data.

The first approaches to concatenative synthesis were motivated and challenged by the high costs for data storage and the limited size of working memory. Therefore, storage-friendly methods such as diphone synthesis or microsegmental synthesis were developed. With these methods, the co-articulatory information between two phones is stored within the units, making the output sound more natural than when using units that only span one individual phone (allophone synthesis, phone synthesis). Diphone synthesis (see e.g. Lenzo and Black, 2000) is based on speech units which start in the middle of one phone and end in the middle of the second one. Microsegmental synthesis (Benzmüller and Barry, 1996a,b) is based on coarticulatorily motivated units on the phone and also sub-phone level, further reducing the demands of storage and computing facilities while enabling more flexible sound concatenation.

Since storage costs decreased, larger data bases have been used. These are e.g. context- and stress-differentiated diphone inventories (e.g. Barry et al., 2001), or, more commonly, units of varying duration, selected according to the demands of the planned utterance (non-uniform unit selection, see e.g. King et al., 1997). To produce an utterance, the best combination of recorded units of different lengths is selected with complex search algorithms. In general, the voices of the unit-selection methods are the ones that people like best. The high degree of naturalness is probably one important reason why concatenative synthesis is the method that is most widely used nowadays for commercial applications, especially as storage becomes available at less and less cost.

The canned, i.e. prerecorded speech is segmented, typically modified, and rejoined (concatenated) to produce smooth utterances of speech. Thus, diphone and unit-selection synthesis techniques also operate on the acoustic level, as does formant synthesis. They use highly advanced search algorithms to select the optimal units and rely on special signal processing techniques to smooth the transitions between units. By using these engineering techniques, an understanding of the underlying speech processes is virtually not necessary any more. The signal manipulation has its limits though. Producing affective utterances from non-affective units for example causes acoustic artifacts that degrade the quality of the synthetic speech, i.e. its naturalness and pleasantness, and perhaps even its intelligibility.

Hidden-Markov-model (HMM) synthesis (e.g. HTS, Tokuda et al., 2000), a variant of the data-driven approach, uses machine-learning techniques to model the properties of speech, and then applies the models to synthesize new speech. However, due to their statistical nature, these models “fail to generate some of the more-interesting and delicate phenomena in speech” (Taylor 2009: 472).

Fundamentally, these techniques are further away from the articulatory level of speech production than formant synthesis. Essentially, big data bases and sophisticated engineering methods have made it possible to develop high quality voices without deeper knowledge of language and speech production. As with formant synthesis, there are, however, hybrid approaches which combine the statistical or HMM methods with articulatorily motivated parameters. This introduces a certain degree of articulatory transparency into the synthesis method, i.e. the values of the parameters can be interpreted in articulatory dimensions (cf. e.g. Picart et al., 2010, 2011, 2012).

2.3.4 Articulatory synthesis

In a broad view, the predecessors of articulatory synthesis are the speaking machines. However, historical mechanical synthesis only imitated human speech mechanisms to a limited extent. Most aspects of the speech production process diverged from natural articulation.

The primary goal was one of acoustic imitation. In contrast, modern articulatory synthesis systems aim to imitate both the speech tract and the movements therein.

In articulatory synthesis, the number of control parameters has increased considerably compared to the traditional hardware synthesis. The handling is more formalized and not as holistically-intuitive as in the historic speaking machine.[3] Moreover, the control mechanisms in articulatory synthesis consist of complex subsystems, again with a great number of different parameters and principles. The musical intuition which was required to 'play' the mechanical synthesizers is replaced by precisely defined elements on a meticulously arranged time line of articulatory movements, i.e. the gestural score (introduced in Section 3.1.5).

Despite the complex internal structure, the main strength and attraction of articulatory synthesis lies in the fact that simple commands to the articulators – and thus changes in the vocal tract shape – suffice to produce the complex patterns of speech. It therefore seems to be the most intuitive way of synthesizing speech. However, this intuitive or transparent method of synthesis needs a large amount of articulatory and aerodynamic-acoustic groundwork. As Carlson (1995: 9932) notes: "Obviously, articulatory synthesis needs considerable understanding of the speech act itself, while models based on coding use such knowledge only to a limited extent."

In their "position paper", Shadle and Damper (2001) have pointed out the attraction of articulatory synthesis, indicating that it is theoretically designated to fulfill all the items on the wish list for an ideal speech synthesizer. The ideal synthesizer should (Shadle and Damper, 2001: 121):

1. "be as intelligible as a human being.
2. sound natural.
3. be able to sound like many different speaker types: male, female; old, young, inbetween; low or high voice.
4. be able to speak in any language.
5. be able to sound like a specific speaker, not just a generic type.
6. be able to sound like an extraordinary speaker, e.g. a singer with a seven-octave voice range, or someone with disordered speech, or an alien with extra sinuses.
7. be able to change to another speaker type, or alter the voice quality of a given speaker, without having to go through as much effort as required for the first voice.
8. have parameter domains that can be conceptualised, so that if it sounds wrong, intuition is useful in fixing it.
9. teach us something and provide opportunities to learn more as we work to produce a commercially usable system."

[3] Holistic would be "Press down the bellows intuitively to generate breathing and different volumes of the voice." Formalized, this is transferred into mathematical curves of pulmonic pressure, serving as input to a computer interface.

This wish list spells out that speech synthesis should on the one hand be an operational reading machine and on the other hand a voice that has personal coloring which can be adjusted with relatively low effort to new demands, once a basic voice has been created. Additionally, Shadle and Damper (2001: 125) conclude that articulatory synthesis has a high "potential" for "extending our knowledge of speech science."

In our view, which complements the one above, current articulatory synthesis is flexible, direct, and, in general, comes without the usual acoustic artifacts known from other types of synthesis. Firstly, it is flexible because once we have established plausible hypotheses of what happens during speech production, articulatory synthesis offers a large degree of freedom to the researcher to manipulate and test the articulatorily transparent parameters one by one. More importantly, we can cover a wide range of speech and paralinguistic phenomena because we can freely synthesize sounds that are unusual in 'regular' speech.

Secondly, it is direct, because due to the articulatory foundation of the synthesizer, certain applications can be handled directly on an articulatory level, such as building new voices, accent morphing or imitating pronunciation pathologies. This is also possible with the black-box-like techniques such as HMM-based speech synthesis (HTS, Tokuda et al., 2000), but in a more indirect way. One has to define acoustic or signal-related equivalents to the articulation beforehand, in order to model the desired auditory impression.

Lastly, with articulatory synthesis, one avoids the common issue of acoustic artifacts, which are generally present e.g. in concatenative synthesis. They often emerge due to boundary mismatches when speech units are concatenated. Since in articulatory synthesis, speech is really produced from scratch, no such acoustic units exist. Therefore the typical issues with these artifacts are not relevant.

The challenge with articulatory synthesis, though, lies in the quality of the acoustic signal (Birkholz, 2006). While other modern kinds of synthesis, working directly on the acoustic level, have optimized the acoustic quality, articulatory synthesis faces challenges on both the acoustic and the articulatory level. Articulatory synthesizers have not yet reached a stable degree of intelligibility and naturalness. Nevertheless, they have been used for phonetic research because naturalness is not the primary criterion there. Articulatory synthesis is regarded as being of "great value to speech researchers" because it "provides a basis for psycho-physical experiments on speech perception" (Hill et al., 1995: 13) where the stimuli can be very precisely controlled regarding articulatory and other factors.

Of the synthesis methods, articulatory synthesis is in general best suited for articulatory investigations. We will present different sub-types of articulatory synthesizers (Section 2.4) before pointing to a perspective in which articulatory synthesis is seen as an articulatory data generator (Section 2.5).

2.4 Types of articulatory synthesizers

While the basic idea of articulatory synthesis is the same in all systems, they vary according to a number of criteria, such as vocal tract model, implementation type, acoustic modeling, and native vs. hybrid systems. This will be discussed in the following.

Vocal tract modeling can e.g. be geometric, physiological (biomechanical), or statistical (following Birkholz, 2006: 17). Geometric models can be constructed from a relatively small set of articulatory data, and they offer high articulatory flexibility because one can freely choose the model parameters to control the vocal tract shape. Biomechanical models mimic biological tissue and can be controlled by muscular activation (cf. e.g. Wilhelms-Tricarico, 1996). Due to their high complexity, they place very high demands on computing power, and to date, the control problem has not been fully solved (Birkholz, 2006: 17). Statistical models are generally constructed from the articulatory data of one particular speaker and thus have a limited number of degrees of freedom. This makes their control easier but limits their flexibility because articulations that were not in the data set can probably not be simulated by the vocal tract model, such as sounds of a foreign language (Birkholz, 2006: 17).

A second distinction among articulatory synthesizers can be made according to implementation type. Some synthesizers are available as a full system, i.e. as one single software program, such as ArtiSynth (Fels et al., 2003, 2005; Vogt et al., 2005), or VTL, while others come as a collection of scripts.

Thirdly, in some systems both articulatory trajectories and acoustic output are integrated in the synthesis process (such as VTL), while others focus on articulatory synthesis primarily in terms of movement output (kinematics). In the latter case, the acoustic module is not permanently implemented within the synthesizer proper (such as TADA at Haskins Laboratories, Nam et al., 2004). Instead, the gestural output has to be forwarded to an acoustic synthesizer, such as formant synthesis.

Similarly, we can distinguish between native articulatory and hybrid or pseudo-articulatory types of synthesis. Native systems, such as VTL, rely on a full articulatory synthesis framework, while hybrid systems combine (pseudo-)articulatory parameters with computationally more efficient acoustic synthesis techniques such as *HLsyn* (cf. e.g. Williams, 1996; Bickley et al., 1997; Hanson et al., 1999).

The articulatory synthesizer *VocalTractLab* (Birkholz, 2006), which is still being extended and improved, is used in this thesis. Regarding the criteria put forward above, it comes as a single complete software program, is a native (full) articulatory synthesizer based on a geometric model of the vocal tract, and offers both articulatory trajectories and acoustic output.

Several other implementations of articulatory synthesizers have been developed e.g. at ICP in Grenoble, France (Badin et al., 2002), at ATR in Kyoto, Japan (Dang and Honda, 2004), at the Institute of Phonetics in Cologne (for German, Kröger, 1998), and at Haskins Laboratories, USA (Rubin et al., 1981) (following Birkholz, 2006: 8).

2.5 A data-acquisition view on articulatory speech synthesis

In this section, we argue that articulatory synthesis can be regarded as a data management framework and articulatory data generator. In this view, articulatory synthesis is closely related to tools for articulatory data acquisition.

All the recording devices described in Section 2.1.3 contribute with their data to an increasingly precise picture of speech articulation. However, the different data sets are not always directly compatible with each other when merging them. One possibility of using the data in a comprehensive way is to integrate them into data management frameworks. Following Guenther et al. (2006), such frameworks can be used for “interpreting and organizing” (p. 28) the huge amounts of collected data. We argue that articulatory speech synthesizers are such frameworks because they provide a shared environment for visualization and organization and also simulation of new data based on previously established foundations.

The human articulatory data help to improve the accuracy of the simulated articulations. VocalTractLab, for instance, incorporates MRI data for its anatomical specification, and functional MRI data of recorded movements are used to define the phone target configurations of the vocal tract (cf. Section 3.1.1). In this sense, the data acquisition tools proper and articulatory synthesis can complement each other: Articulatory synthesis which is based on extensive data collections can provide a realistic virtual speaker that can generate tailored, realistic articulatory data for speech articulation research, even for subtle details of articulation.

As an additional feature, the acoustic output that corresponds to the simulated articulatory movements is generated as well. It is therefore possible to verify new articulation patterns via evaluation of their corresponding acoustics. This is our motivation to use articulatory speech synthesis and not to record and analyze original articulatory data in the traditional manner. These two approaches are delineated in the following.

In the traditional phonetic experiment, we have speakers produce certain utterances under the assumption that they will do it ‘correctly’. Auditory control is used to detect clear, i.e. audible divergences. During the recordings, a – due to the experimental setup – restricted number of articulatory parameters is captured along with the acoustic signal. The parameters are then subjected to articulatory analysis. The decisions on which speech material and which articulatory parameters to record are guided by their assumed relevance to the utterance, i.e. by phonetic knowledge and phonetic hypotheses. As many participants as is justifiable for the given method are recorded to establish a plausible chain of cause and effect from articulation to audible properties.

When we use articulatory synthesis for data acquisition, we have an unambiguous model-specific causal chain because we employ one and the same system to generate the articulation and its audible consequences. As it is with the traditional phonetic experiment, the phonetic knowledge is put at the beginning of the experiment to determine the utterances and the relevant articulatory parameters. After the synthetic (simulated) data collection, the utterances are, as in the traditional manner, controlled auditorily or judged in a perception test. In this sense VTL can serve as a tool for data acquisition. But in contrast to the traditional phonetic experiment, it generates in one pass articulatory and acoustic data, i.e. VTL replaces the speakers, the digitization and the articulatory sensors.

To conclude, we review the relevant attributes of tools for data acquisition again (cf. Table 2.1) for an articulatory (synthesis) framework which integrates all relevant data acquired from humans: By using a sophisticated control concept to produce sensible articulatory movements, the simulated movement data can cover the whole vocal tract and not just parts of it, have a very high time resolution and a very high spatial resolution. The data are fully three-dimensional, their acquisition (i.e. synthesis) involves no health risk, the tool is portable (piece of software), and is, for the end user, of low cost.

As has been laid out in this chapter, articulatory synthesis seems particularly suited to complement traditional data acquisition tools. We will therefore use it for articulatory research in this thesis, evaluating how well it deals with questions regarding fine articulatory detail.

Chapter 3

The articulatory speech synthesizer VocalTractLab

This chapter provides background information about the articulatory speech synthesizer VocalTractLab (VTL) that helps to understand the technical details and terminology mentioned throughout the empirical part of the thesis. Section 3.1 describes the main modules of the synthesizer, interwoven with *basic* descriptions of the speech articulation processes in humans,[1] pointing to relevant differences between human speech articulation and its artificial analog in VTL. This encompasses a presentation of the concepts of articulatory gestures and vocal tract targets, and the introduction of related terminology. Information on relevant fine articulatory details will be presented in the introductory sections of each of the experiment chapters. Section 3.2 presents information about the suitability of VTL for our experiments but also foresees a number of limitations which influence our experimental design and which have to be taken into account during data interpretation. Finally, in Section 3.3 we briefly list further applications of VTL, other than articulatory-perceptual research.

Speaking is a very complex process that is by no means limited to the mere generation of acoustic output. It is a process of communicative action, and offers numerous perspectives on how it operates. In this chapter, we focus on the *production* aspects of the speech chain, i.e. leaving out for the moment the acoustic and perceptual aspects. (Some aspects of acoustics and perception are discussed in Chapter 4.) We will limit our perspective to the 'mechanical' generation of speech and the aerodynamic and spatial prerequisites for it. In other words, speech generation in this view is based on the idea that we assume some air flow as energy source, which is modified into audible energy by phonation, and the use of vocal organs modifies the acoustic source signal to produce the distinct sounds of a language which are radiated from the lips and the nose.

[1] The presentation of basic human speech articulation processes goes along the lines of Crystal (1997), Chapter 22. A more extensive account can be found e.g. in Hardcastle (1976).

A special focus is put on physiological and production aspects that are *not* implemented in VTL, to point out the resulting differences in the capabilities of the machine vs. the human. This includes e.g. speech breathing and the production of clicks. Additionally, it is important to note that the planning and integration of the single movements into one coherent complex speech movement is only possible through the central control in the brain. However, our focus lies on the manifested speech movements, and aspects of the brain, nerves, and muscles are not discussed in depth here. Finally, we omit discussions of feedback mechanisms. These are necessary for humans to acquire speech and maintain intelligible pronunciation, and include auditory and proprioceptive feedback. They are not implemented in this version of the synthesizer, although extensions have been put forward which simulate speech acquisition and feedback mechanisms (Birkholz and Kröger, 2007, Kröger et al., 2006, Kröger et al., 2007).

The general important strength of VTL, or articulatory synthesis, is located in the following aspect. *Qualitative* descriptions, such as ‘lip closure’, are mapped onto VTL vocal tract specifications, and thereby become *quantitative* statements. In other words, VTL provides a level of spatial description that is very fine grained and specific to an individual vocal tract. This transformation from the qualitative level to the quantitative level makes it possible to work with fine articulatory detail because minute differences in geometry can be modeled reliably. Depending on their nature and location, they may have a relevant impact on the acoustic output. The fine articulatory details of the actual movements are of interest because they are the key information to our paralinguistic phenomena: Subtle variations in the manner of speaking are often all we need to produce this information.

3.1 Overview of main components

For the simulation of articulatory processes and the generation of the corresponding speech waves by an articulatory synthesizer, a number of different components and mechanisms are necessary. The main modules of the synthesizer VocalTractLab are described in this section, concentrating on a perspective that is rather user-oriented, and mainly following the descriptions found in Kröger and Birkholz (2007). The main components are: the supraglottal system, the dominance model for coarticulation, the glottis, the subglottal system, the gestural control concept and the acoustic simulation. This account is intended primarily as a reference and introduction to the terminology used later during the empirical part of the thesis when referring to the synthesis technical details. Thus for further reference and more information on technical details, please see Kröger and Birkholz (2007) (general overview), Birkholz (2006) (whole system in detail), Birkholz et al. (2006); Birkholz and Kröger (2006) (dominance model and vocal tract modeling), Birkholz et al. (2011a); Birkholz (2007b); Kröger and Birkholz (2007) (articulation, gestural control), Birkholz et al. (2007a); Birkholz and Jackèl (2004) (acoustic simulation). Information about the software and the principles it is based on can also be found in an extended manual (Birkholz, 2013c).

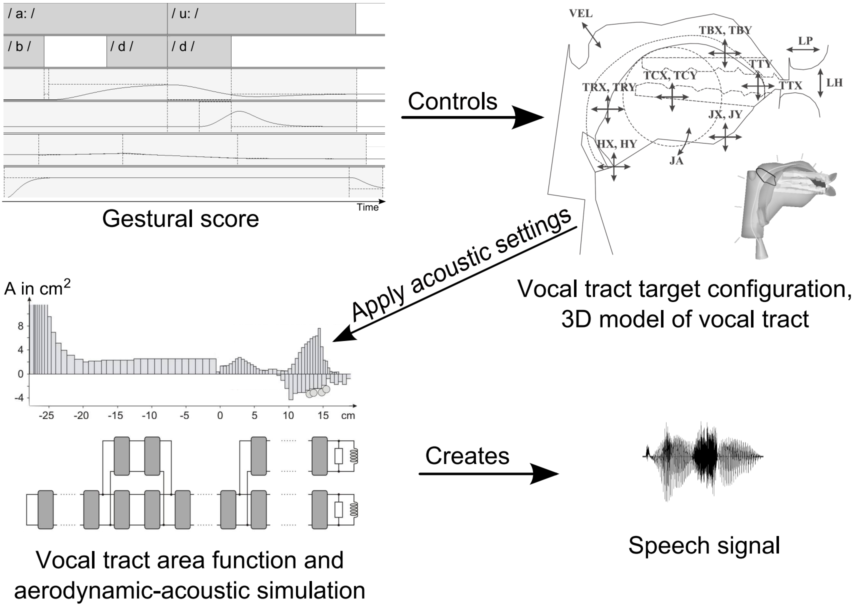
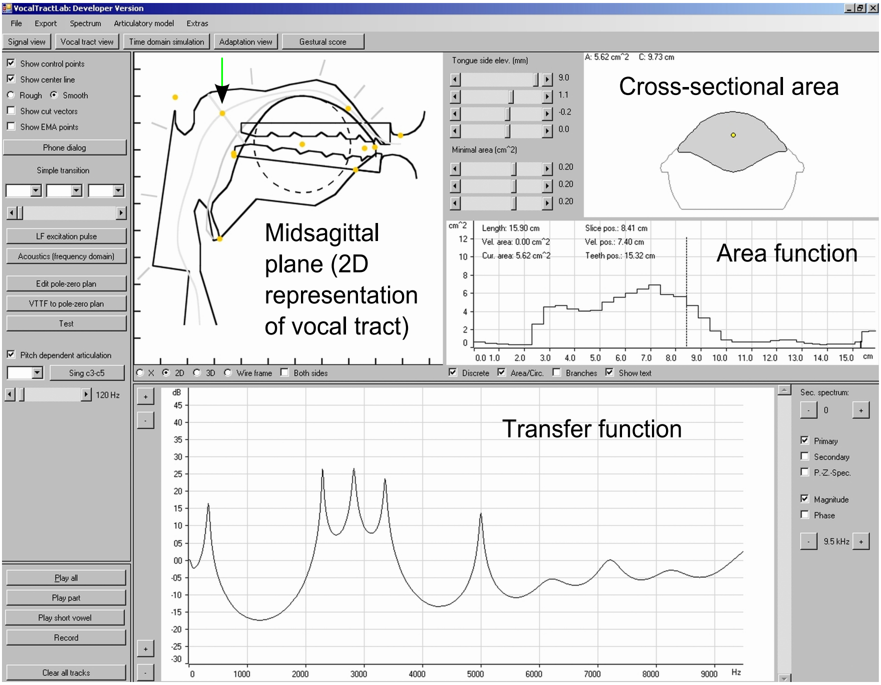


Figure 3.1: Overview of the synthesis process in VocalTractLab. Details see Section 3.1.

In a nutshell, the process to obtain speech output in this synthesizer is the following (see also Figure 3.1): The basis for synthesizing virtual speech movements is located in the representation of the anatomy of the current speaker (supraglottal system), and a means to control the movements within the given vocal tract (gestural control concept). By applying predefined acoustic settings, the speech movements are rendered into precisely defined articulatory trajectories. Every movement (or trajectory) causes a deformation of the vocal tract shape, resulting in a specific vocal tract area function, and thus determines the transfer function of the vocal apparatus. From this, combined with an appropriate source signal, the software calculates the final speech wave during aerodynamic-acoustic simulation, thus creating the link between the articulatory and the acoustic domain. Details of the anatomy, control, and simulation are described in the following sections.

The synthesizer provides a graphical user interface (GUI) to enable manual adjustment of parameters, to monitor their effects, such as variations in the transfer function, and to execute the synthesis functionalities. A detailed overview of the different windows of the GUI, including screenshots, can be found in the Appendix of Birkholz (2006). Figure 3.2 shows the “vocal tract view”, which can be used to create new sound configurations by hand.

The software can also be accessed by a programming interface (API), which can be used to externally control the synthesizer and obtain articulatory or acoustic data from it. We used this interface in the experiment on vocal aging (Chapter 7).



☑ Show control points
☑ Show center line
○ Rough ● Smooth
□ Show cut vectors
□ Show EMA points
☑ Pitch dependent articulation
120 Hz
○ X ● 2D ○ 3D ○ Wire frame □ Both sides

Tongue side elev. (mm): 9.0, 1.1, -0.2, 0.0
Minimal area (cm^2): 0.20, 0.20, 0.20

A: 5.62 cm^2 C: 9.73 cm

Length: 15.90 cm
Vel. area: 0.00 cm^2
Cur. area: 5.62 cm^2
Slice pos.: 8.41 cm
Vel. pos.: 7.40 cm
Teeth pos.: 15.32 cm

☑ Discrete ☑ Area/Circ. □ Branches ☑ Show text

Sec. spectrum: 0
☑ Primary
□ Secondary
□ P.-Z.-Spec.
☑ Magnitude
□ Phase
9.5 kHz

Figure 3.2: "Vocal Tract View" of the synthesizer, showing details for the vowel [eː]. The gray dots in the **2D vocal tract contour** (upper left) allow manual adjustments of the parameters by dragging the points within their allowed ranges. One gray dot is fixed to the center line (gray cross of lines, where the arrow points). The corresponding shape of the perpendicular **cross-sectional area** at that point on the center line is shown on the upper right. (A 3D version of this sample cross-sectional shape is shown in Figure 3.3, left side.) The set of horizontal slide bars to the right to the vocal tract allow the adjustment of the three-dimensional parameters of the vocal tract, which control the tongue side elevation and the minimal cross sectional areas at three points in the vocal tract. These minimal areas are important because they prevent complete closure of the vocal tract where we only want fricative approximations when designing new sounds manually. Below the slide bars, the current spatial dimensions of the speech apparatus are displayed in terms of a discrete **area function** along the center line through the speech apparatus, ranging from the lungs to the lips. The nasal branch is not displayed here. At the bottom, the **transfer function** of the current vocal tract configuration is shown.

3.1.1 The supraglottal system

The human supraglottal system, or vocal tract, is probably one of the most versatile parts of the speech apparatus. Moving the mobile vocal organs to produce speech sounds is called articulation, and the shape of the resulting cavities (pharyngeal, oral, nasal) determine by their resonance behavior the characteristics of these sounds. Typical active articulators are the tongue, soft palate, velum, and jaw. Passive articulators are the teeth, the alveolar ridge, and the hard palate. Finally, the pharynx can be seen as a more or less active articulator.

VTL uses a geometric model of the vocal tract, which is defined as a cavity. Its shape is controlled by a set of vocal tract (VT) parameters. They are designed to have only local influence on the overall geometric shape. The main differences to a human vocal tract are sketched, following Crystal (1997), before going into more technical details about the vocal tract model. After that, we will discuss some general properties of this model.

Human-machine comparison

In humans, the pharyngeal tube can be narrowed or widened both to produce pharyngeal fricatives such as [ħ ʕ] and to modify other sounds by pharyngealization. In VocalTractLab, the pharyngeal wall is a passive articulator but the pharyngeal tube can be modified indirectly by rearranging the back part of the tongue to influence the width of the tube. The human soft palate or velum consists of muscular tissue including the uvula. In VocalTractLab, this is represented accordingly with a movable contour at the roof of the mouth. It changes into the hard palate, the alveolar ridge and finally the teeth. They all are immovable contours representing passive articulators.

The human lips are mainly made up of the mouth-encircling muscle *orbicularis oris*, which combines with the extrinsic lateral muscles to control lip movement. The lips can be closed or held apart in varying degrees. They participate in many sounds, providing friction constrictions in [ɸ β], and rounding, protrusion or spreading in e.g. vowels or smiled speech. In VTL, the lips are not associated with a muscle, instead only their resulting contour is modeled and can produce all the above-mentioned kinds of articulations.

The lower jaw or mandible bone in humans is in itself inflexible but its position can be changed as a whole, influencing the vertical distance between the teeth and often also the position of the lips as well as the tongue. In VTL, the jaw angle can be changed realistically, but this has virtually no influence on the position of the tongue. Furthermore, although jaw angle manipulations do influence the position of the lower lip, the upper lip is shifted automatically in a way that the vertical lip distance is kept constant.

The human tongue is essentially a muscle complex that can move in all directions by means of extrinsic muscle actions, and in any of the achieved positions can additionally be shaped by intrinsic muscles. This makes it very versatile. While there are no obvious anatomical sections in the tongue, it is classified relative to the roof of the mouth into dorsum, blade, apex, and rims. In VTL, the tongue is represented only by its surface contour, including a connection from the slightly thinned out apex to the floor of the mouth. Similar to the human sectioning, it is divided into adjacent sections, each shaped by local control variables. Additionally, the shape of the rims can be modeled by adjusting the parameters of tongue side elevation ($\mathrm{TS}_{1..4}$, cf. p. 30).

Features of the vocal tract model

After having sketched main differences between VTL and the human vocal tract, we describe the model vocal tract in more detail. Summing up the above, the model defines the surfaces of both the articulators, such as the tongue, the lips, or the uvula, and their more rigid counterparts, such as all walls of the vocal tract and the teeth. While the vocal tract walls are modeled all the way down to the larynx including the epiglottis, the glottis itself is not part of this vocal tract model. A separate model of the vocal folds, which generate the source signal in the simulations, is described in Section 3.1.3. Additional parts of the vocal system include the lungs, trachea, and nasal cavities (cf. Section 3.1.6).

The geometry of the oral and pharyngeal cavity is modeled by a set of wireframe meshes (cf. Kröger and Birkholz, 2007) and can be rendered into a 3D view of the vocal tract as shown in Figure 3.3 (left side). Different surface characteristics, such as absorbing walls, are also specified. The specific contour of the rigid parts of the vocal tract is derived from scans of one adult male German native speaker (Birkholz and Kröger, 2006). Thus it is his anatomy that is implemented in this vocal tract shape. Additionally, the same speaker was recorded with Magnetic Resonance Imaging (MRI, cf. also Section 2.1.3) during speech production to record *default phone configurations* to build a *standard phone set* for German.

For each speech sound, midsagittal vocal tract contours were traced in the recorded images. The outlines of the model vocal tract were then adjusted manually to closely match the tracings of the recordings. Finally, the contour for each sound was stored using the vocal tract parameters introduced below. Thus, each speech sound can be activated by retrieving its specific, predefined parameter setting. This setting for a single sound is called the *vocal tract target configuration* for that sound. It represents sound definitions on the *quantitative* level, i.e. the spatially precisely defined basis for speech sound articulation in a given vocal tract, based on requirements on the *qualitative* or symbolic level, i.e. phone labels.

Each sound definition consists of a list of *vocal tract (VT) parameters*. They define the position, orientation and shape of the different structures of the vocal tract. They are stored in an XML configuration file, the *speaker configuration file*, and can be accessed by opening the *phone dialog* in the GUI. Their names and functions are listed in Table 3.1. Figure 3.3 (right side) depicts the position and directions of influence of the most important parameters, which are also described in the following.

LP and LH determine the degree of protrusion of the lips and the vertical distance between them. They are e.g. used for smiled speech control (Chapter 6). The shape of the tongue in the midsagittal plane is defined by four pairs of parameters, describing the position of different parts of the tongue. Tongue tip (TTX, TTY) and tongue center (TCX, TCY) are

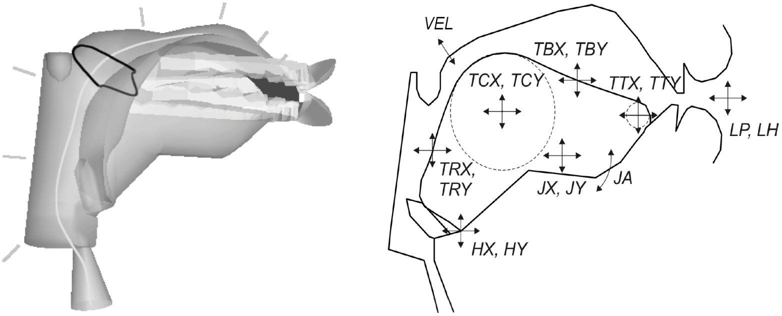


Figure 3.3: **Left:** 3D view of the vocal tract for the vowel [eː], showing the *center line* (white line) at which we can visualize the local cross-sectional area. **Right:** Vocal tract parameters on the midsagittal plane and their directions of influence (arrows), adapted from Kröger and Birkholz (2007: 177), with friendly permission by the author.

Table 3.1: Description of the vocal tract parameters, as they can be found in the phone dialog of the VTL GUI and similarly in Birkholz (2006: Table 2.2). They are stored in the speaker configuration file. Each sound is defined by a complete list of these parameters. Pairs of labels define x,y-coordinates in an underlying Cartesian coordinate system.

Label	**Range**	**Neutral**	**Unit**	**Description**
HX	[0.0, 1.0]	1.00		Measurement for the position of the hyoid bone
HY	[-6.0, -3.5]	-4.75	cm	Vertical position of the hyoid bone
JX	[-0.5, 0.5]	0.00	cm	Position of the lower jaw
JY	[-1.8, -1.2]	-1.50	cm	
JA	[-0.2, 0.0]	-0.10	rad	Opening angle of the lower jaw
LP	[-1.0, 1.0]	-0.07		Measurement for the degree of lip protrusion
LH	[-0.5, 4.0]	0.95	cm	Vertical distance between the upper and lower lips
VEL	[0.0, 1.0]	0.00		Measurement for the position of the velum
TCX	[-3.0, 4.0]	-0.40	cm	Position of the tongue center (tongue body)
TCY	[-3.0, 1.5]	-1.46	cm	
TCRX	[1.0, 2.0]	1.80	cm	Tongue center radius
TCRY	[1.0, 2.0]	1.80	cm	
TTX	[1.5, 5.5]	4.07	cm	Position of the tongue tip
TTY	[-3.0, 1.5]	-1.88	cm	
TBX	[-3.0, 4.0]	2.00	cm	Position of the tongue blade
TBY	[-3.0, 5.0]	0.50	cm	
TRX	[-4.0, 2.0]	0.00	cm	Position of the tongue root
TRY	[-6.0, 0.0]	0.00	cm	
TS1	[-1.4, 1.0]	0.00	cm	Tongue side elevation at the root
TS2	[-1.4, 1.0]	0.06	cm	Tongue side elevation at the dorsum
TS3	[-1.4, 1.0]	0.15	cm	Tongue side elevation at the blade
TS4	[-1.4, 1.0]	0.15	cm	Tongue side elevation at the tip
MA1	[0.0, 0.3]	0.00	cm^2	Minimal cross sectional area behind the tongue tip
MA2	[0.0, 0.3]	0.00	cm^2	Minimal cross sectional area at the tongue tip
MA3	[0.0, 0.3]	0.00	cm^2	Minimal cross sectional area in front of the tongue tip

represented by circular arcs, and the parameters represent their center positions (Kröger and Birkholz, 2007). Tongue body (TBX, TBY) and tongue root (TRX, TRY) are represented by two second order Bézier curves, and the parameters are the Cartesian coordinates of their medians points. Since the vocal tract model is three-dimensional, additional parameters define the state of non-midsagittal parts of the tongue: The relative height of the tongue sides is provided by four additional tongue parameters ($\mathrm{TS}_{1..4}$). Minimal cross-sectional areas at, before and behind the tongue tip are given by three minimal-area parameters ($\mathrm{MA}_{1..3}$) to avoid unwanted articulatory closures e.g. when defining fricatives.

Besides lips and tongue, the other important vocal tract parameters cover the position of the velum, the lower jaw, and the hyoid bone (connected with the larynx). The velum can be lowered or raised in varying degrees, specified by VEL. Depending on its position, the velum automatically changes its shape, e.g. when being pressed upwards against the vocal tract walls. The position of the jaw is provided in two ways, firstly by defining its general position in the coordinate system (JX, JY), and secondly by specifying a degree of rotation (JA) which represents the jaw opening angle. The final pair of parameters (HX, HY) specify the position of the hyoid bone. Due to its tight link to the larynx, which is based on human physiology, these parameters also specify the position of the larynx, and thus serve to control larynx height e.g. in smiled vowels (Chapter 6).

The parameters introduced here are used repeatedly in the empirical part of the thesis. By systematically varying their actual values, we induce slight changes in the vocal tract shape which in turn produce different acoustic output. As described above, the parameters can be manually manipulated in the GUI. However, for vowels it is also possible to use a specific functionality that optimizes the vocal tract shape to obtain a given formant structure. This so-called formant optimization algorithm is applied in the regional accent experiment (Chapter 10) and is explained in Section 4.2.4.

While the oral cavity is fully modeled geometrically, the nasal system is represented only as a one-dimensional tube model, providing the area function of the nasal cavity and the para-nasal sinuses. This suffices to determine the geometric foundation for the acoustic simulation (see Section 3.1.6).

The nature of the vocal tract model

As implied in the human-machine comparison (p. 27), the vocal tract model does not fully cover prevalent natural co-dependencies between certain articulators such as the connection between the lower jaw and tongue height. This will be taken into account when analyzing related data in the empirical part of the thesis. On the other hand, connections such as between the hyoid and the larynx *are* modeled, so that for specifying larynx height we can use the vertical component of the position of the hyoid bone (HY).

The reasons for using a model of the vocal tract that is designed largely without explicit biomechanical dependencies among its individual parts are portrayed in the following (Birkholz 2012, pers. comm.), because we often mention that it would be helpful if they were indeed modeled. But there is a reasonable motivation for not implementing them at the basis of the geometric vocal tract setup.

Firstly, the geometric shape of the cavity is the central determinator of the acoustic output. This output in turn represents the mental target space for pronunciation, according to current research, since we want to produce acoustics (in order to be understood) and not only articulatory movements per se (cf. e.g. Nieto-Castanon et al., 2005).

Secondly, the current system is ready to be extended any time by introducing an additional layer of parameters which model biomechanical dependencies. However, the basic geometric shape of the cavity is always needed because it is the basis for acoustic simulation.

Thirdly, the main dependencies such as those between jaw and tongue or tongue center and tongue tip can be produced only with an adequate biomechanical model. However, these models are not yet sufficiently evaluated. For a single muscle one needs a lot of parameters. But their precise relationships are only sparsely documented and biomechanical models often rely on only small sets of empirical data and many assumptions. As a consequence, models of muscles have to be considerably simplified, and interdependencies between them can only be considered in a rudimentary way, if at all. Since many basic details have to rely on estimated values, variation in resulting muscle shape has to rely on estimated muscle deformations from estimated forces. This leads to estimated cavity shapes which are not constructive for a fully functional articulatory speech synthesizer that aims at producing reliable acoustic output.

Lastly, the human speaker does have the capability to control the jaw and the tongue independently of each other. This is why strict dependency constraints perhaps do not always make sense. Of course, not all kinds of movements should be allowed and possible, but these restrictions can also be defined in a geometric model, namely by providing boundary conditions (values) which prevent gross mispositioning of the articulators.

Since the geometric model only describes the shape of the surfaces of the articulators, it is not capable of handling volume constancy. However, not even biomechanical models always manage to keep the volume of an articulator, such as the tongue, constant. Perhaps this is not even necessary because, especially in the case of the tongue, the elastic mouth floor bends with tongue movement. Thus, modeling articulator volume interferes with varying total volumes of the vocal tract anyway.

3.1.2 Dominance model for coarticulation

In the section above, we introduced the parameters that determine the shape of the vocal tract when a sound is uttered by itself, such as an isolated stationary vowel. We now discuss what happens when two or more sounds are articulated in sequence. It will always be the case that one articulator is occupied to reach the spatial target for a sound A, while another sound B also influences this articulator to move to a different location (coarticulation). Thus, targets are not fully reached due to the influence of neighboring sounds. Therefore, the vocal tract parameter configurations for the sounds can be understood as complete but theoretical target configurations. The targets are theoretically fully specified to provide a point in space which serves as a goal for the articulatory movements. But only in their linguistically relevant parts they have to be executed for proper segmental articulation.

To resolve the coarticulation conflicts, a dominance model is put into place (cf. also Birkholz et al., 2006: 875f). It determines the final trajectory of an articulator as a calculated compromise according to dominance values given in the phone definitions (speaker configuration file). For this to work, every vocal tract parameter that was introduced above not only has its spatial definition but also a dominance value. Its range varies between 0 (minimally dominant) and 1 (maximally dominant) and determines the salience of a vocal tract parameter for the production of that sound, or the actual linguistically relevant articulation aspects of each sound.

For a [p] e.g., bilabial closure (LH = 0) is essential, thus the dominance value of LH, the parameter for the lip distance, should be maximal (dominance = 1, corresponding to 100%) to ensure that the closure is reached in any case. Usually, the articulation of consonants allows for certain degrees of freedom, whereas the definitions of vowels are more strict. In the default speaker file, all vocal tract parameters for vowels have a dominance of 100%. In contrast, in a consonant such as [p] it is not essential for the tongue body to be in a certain position since the lips acoustically obscure the oral cavity behind the labial closure. So the dominance values for the tongue body are set to minimum for [p].

For certain coarticulation processes, it is important to set some dominance values to sensible intermediate values. For instance, velarization of [l] into ‘dark l’, depending on the vowel context, will take place if the tongue body dominance values are rather low. For German, this kind of coarticulation is not desired, so the corresponding dominance values should be somewhat higher. If they are too high though, the [l] will sound hyperarticulated.

Individually adjusted dominance values are stored in the default speaker file, which is provided within the framework of VTL. In our experiments, however, we found that some sounds needed some more ‘tweaking’ in order to show appropriate coarticulatory behavior when synthesizing words (cf. Chapter 11).

3.1.3 Glottis

Humans are capable of producing a variety of different phonatory settings, using the complex structure of the larynx (cf. e.g. Laver, 1980). This tube in the upper part of the trachea consists of ligaments, membranes and cartilages, most importantly the thyroid, cricoid and arytenoid cartilages. The glottis itself is the gap between the muscular tissue of the vocal folds in the center of the larynx. Above the vocal folds we find the ventricular, or false vocal folds.

The true vocal folds are commonly used in speech for phonation, they determine the fundamental frequency, and depending on the inner state of the vocal folds (tension) and the

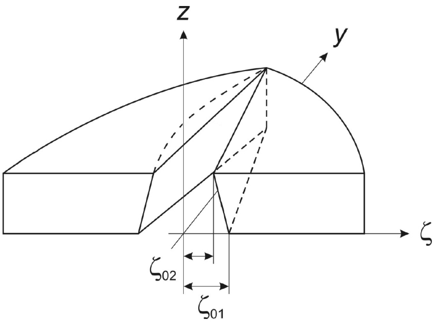


Figure 3.4: Model of the vocal folds. Illustration adapted from Kröger and Birkholz (2007: 179). With friendly permission by the author. List of parameters is given in Table 3.2.

Table 3.2: Description of the glottal parameters, as can be found in the glottis dialog of the VTL GUI (cf. also Birkholz, 2006: 41, Table 2.5). An illustration is presented in Figure 3.4.

Label	Description	Value range
Fundamental frequency	f_0	50 – 400 Hz
Subglottal (pulmonic) pressure	p_{sub}	0 – 1200 Pa
Degree of abduction at lower edge of vocal folds	ζ_{01}	-0.5 – 3 mm
Degree of abduction at upper edge of vocal folds	ζ_{02}	-0.5 – 3 mm
Phase difference between upper and lower edge (vertical phase difference, phase lag)	ϕ	0 – π rad
Width of the active opening of the posterior chink (glottal leak)	ΔA_{chink}	-5 – 5 mm^2

cartilages around them we can produce different voice qualities during speech. The false vocal folds are not typically used for speaking, unless certain voice qualities are desired, in some singing or chant styles (e.g. Bailly et al., 2010; Esling, 2002), and in the case of voice disorders.

The complex human anatomy and the variety of possible configurations within the larynx is simulated in VTL in a somewhat simplified and functional manner. The state of the glottis is defined by a parametric model which describes the surface geometry of the vocal folds. The model is based on the model by Titze (1984) and has been slightly adapted (cf. Kröger and Birkholz, 2007, Birkholz, 2006: 39ff). The model generates cross-sectional areas at the lower and upper end of the glottis. They are mapped on tube sections, and in this way the glottis' spatial structure becomes part of the overall tube system that is used for the acoustic simulation (Kröger and Birkholz, 2007, cf. Section 3.1.6). An illustration of the model is shown in Figure 3.4, its parameters are listed in Table 3.2. The false vocal folds are not modeled.

The basic function of the glottis model is to provide a source signal for the voiced excitation in the synthesizer, with a specific fundamental frequency. Since the model represents a simplified version of the human glottis, the variety and the degree of detail in the settings is restricted. The most important parameters are abduction, adduction and f_0. We use the parameters e.g. to manipulate the voice quality of our synthesized samples in voices that are intended to simulate different speaker ages (Chapter 7).

Although the glottis component seems disjunct from the supraglottal vocal tract, it should be noted that the different parts interact with each other acoustically. To obtain a natural sounding voice, vocal tract characteristics and fundamental frequency, or larynx size, should match. A current limitation of the synthesizer is the fact that the excitation signal sounds more natural for male than for female voices due to the nature of the vocal fold model. Additionally, the resonance characteristics of the vocal tract match those of male voices better because the synthesizer's vocal tract was adapted to the anatomy of an adult *male* speaker (Birkholz and Kröger, 2006). This makes the resonance characteristics sound slightly better when combined with a male (low f_0) voice excitation than with the f_0 ranges of women or children.

3.1.4 Subglottal system

The subglottal system encompasses those parts of the speech apparatus that are located below the human larynx. Its largest part are the lungs, embedded in the thoracic cage, bounded by the diaphragm, the sternum and ribs, and the spinal column. The upper end of the lungs leads to the trachea which ends at the larynx.

The main function of the lungs in speech is to produce an outgoing (egressive) air flow at a relatively constant pressure level. While the routine usage of the lungs for speech is to provide egressive pulmonic air flow, humans are also capable of talking on an ingressive air flow. This mode contributes to the versatility of the human voice and its naturalness and is allegedly used for 'exceptions'. It can e.g. be found when people count "under their breath" (Crystal, 1997: 125), i.e. rapidly for a longer time without pausing to inhale; when someone laughs and tries to speak at the same time; or when a speaker is out of breath and nevertheless attempts to talk. A widespread application of ingressive voice however seems to take place during backchanneling (Eklund, 2007). Furthermore, humans can produce sounds both with and without using the mechanisms of the lungs, enabling the production of clicks, ejectives, and implosives. VTL however only works with simulated air that is coming from the lungs. It is therefore limited to egressive pulmonic sounds.

Another difference is that egressive air flow is only used in isolation to provide the energy source for audible speech. It is not in any way part of simulating a respiratory cycle. From humans we know (see e.g. Reetz, 2003: 108, Conrad and Schönle, 1979) that in normal breathing we take a roughly equally long time to inhale and to exhale. When speaking, we switch to a mode called speech breathing, in which we inhale quickly (taking about 10 % of the time in a cycle) and maintain a rather stable level of subglottal air pressure while slowly releasing the air when producing the speech sounds. Compared to this cycle, VTL selectively simulates the exhalation phase of the speech breathing mode. With some articulatory tweaking in the VTL pharynx, it is possible to make breathing noise audible (as done in Chapter 8). In this way, it can function as a non-verbal sign of communication.

From the technical perspective, the subglottal system is, as is the nasal system, represented only as an area function, omitting a full geometric model. The most important subglottal parameter in our experiments is the pulmonic pressure. The pressure level can be set to values from 0 to 1200 Pa in the synthesizer's GUI. When synthesizing 'regular' speech, good results are obtained when pulmonic pressure stays rather constant and declines towards the end of a phrase. The situation is different when synthesizing laughter. Here, an elaborate use of varying pressure levels is necessary (cf. Chapter 8).

3.1.5 Gestural control concept and its components

We need a time line as a means to plan and define an utterance, be it for actual fluent speech or any well-formed sequence of sounds or even merely a single sound such as a plosive. For humans, it is obvious that the planning step is executed in the brain (cf. e.g. Levelt et al., 1999). However, since VocalTractLab does not work with a linkage to an artificial brain, we focus on the actual generation of the articulatory movements in space. Even if the brain structures are neglected from the description, it still remains a complex task to define and generate the speech movements in the model.

The method that is used in VTL is inspired by articulatory phonology (Browman and Goldstein, 1992) and includes a *gestural score*, which consists of different tracks or tiers, filled with individual *gestures* as the basic unit of temporal control. In the following, we sketch the two components, i.e. the gesture and the score, before describing the actual machinery behind these relatively abstract descriptions: the gestural control concept.

Gestures

The general concept of articulatory gestures and articulatory phonology, as e.g. discussed in Browman and Goldstein (1992), Fowler and Saltzman (1993) and Kröger (1998), refers to "linguistically significant actions of structures of the vocal tract" (Fowler and Saltzman, 1993: 172). The variability of speech, as observed on the acoustic surface, can be explained by basic attributes of articulatory gestures: amplitude, duration, velocity, and temporal overlap with other gestures. This simplicity constitutes the attractiveness of articulatory phonology. An adaptation of these concepts (see also Birkholz, 2006: 138), based on the target approximation model of Xu and Wang (2001), has been implemented in the synthesizer.

In VTL, a gesture is used in a broad sense in that it currently does not primarily denote individual phonologically significant gestures, such as 'lip closure', but it refers to complete spatial descriptions based on the vocal tract target configurations introduced in Section 3.1.1. Furthermore, in this thesis (and in VTL), items are generally called gestures if they are items on the gestural score, even if they refer to fundamental frequency or subglottal pressure contours, and not only to the supraglottal area.

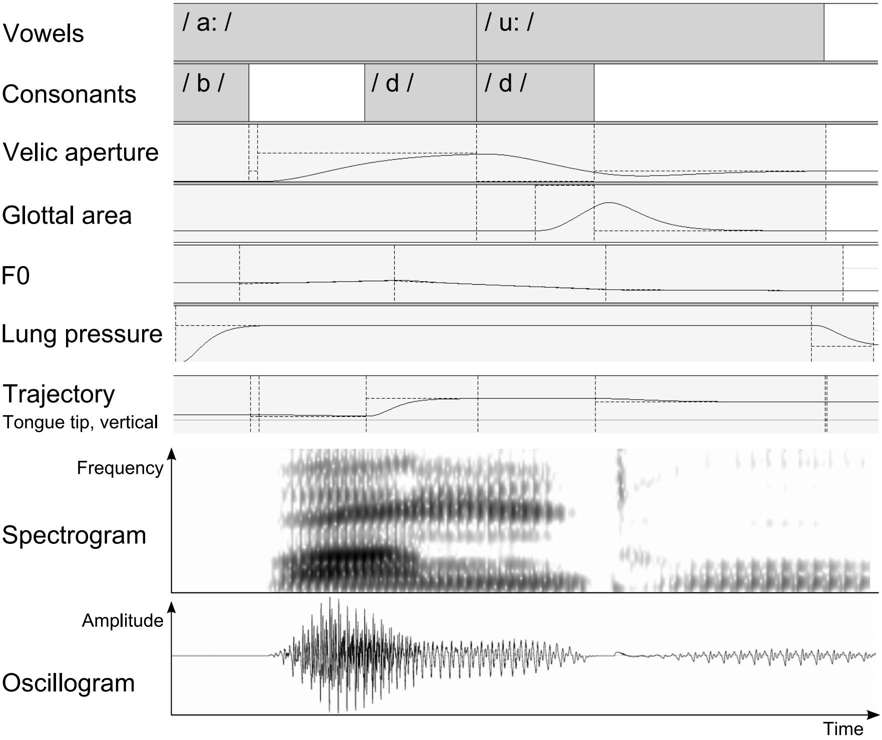


Figure 3.5: Top six tiers: Sample gestural score depicting a possible way of articulating the word <Bantu>. Below: Resulting vertical position of the tongue tip (TTY) over time, exemplifying that resulting trajectories of single vocal tract parameters can also be visualized in alignment with the gestural score. Bottom: Corresponding spectrogram and oscillogram are added to illustrate the acoustic output of the articulatory processes.

The gestures have distinct durations and amplitudes, and they can vary in the speed with which a target is approached (articulatory velocity, articulatory effort, vocal effort, or slope of onset). If the duration of a gesture is too short to reach the desired target completely, the articulatory phenomenon of hypo-articulation occurs (cf. Lindblom, 1990). This can also result in merged or deleted sounds, and resembles articulatory reductions which are typical of everyday speech (cf. also Birkholz, 2007b; Browman and Goldstein, 1989, 1990).

The strength of using gestures as defined in VTL is that they help modularize sound production. For different sounds which share the same place of articulation, the same supra-glottal gestures can be invoked, combined with varying velic or glottal gestures, as will be explained in the next section.

Gestural score

The gestural score is a means of controlling the artificial speech tract over time by specifying the temporal coordination of all necessary articulatory gestures (gestural alignment). Each functional group of gestures is represented on a different *tier* on the gestural score, as shown in Figure 3.5.

For segmental articulation, the most important tiers are the consonantal and the vocalic tier, complemented by the velic and glottal tiers. By assigning a phone label to a supraglottal gesture, the vocal tract parameter configuration of that sound is retrieved from the speaker configuration file (cf. Section 3.1.4, p. 28), and gradually approached by the model articulators. When no gesture is specified on a tier, a default gesture may be assumed, such as the Schwa if no vowel is set on the vocalic tier. On the consonantal tier, no gesture is executed at all when no label is specified.

The velic and glottal tiers are used to specify properties of nasals or aspirated plosives. The definitions for these sounds regarding the *oral* cavity are taken from their homorganic unaspirated plosives, see e.g. the usage of /d/ in Figure 3.5 appearing as [n] or [t] in the word <Bantu>. Simply by adding aspiration for the required delayed adduction of the vocal folds after the stop release, the voicing automatically disappears, and a voiceless aspirated plosive is articulated ([t]). This is achieved by placing an abduction gesture on the glottal tier. Similarly, when nasal coupling is added, the voiced plosive is changed into its homorganic nasal ([n]). This is done by placing an "open" gesture for velic aperture on the velic tier.

For supra-segmental articulation, the relevant tiers include the pulmonic tier and the f_0 tier. The pulmonic values are handled like gestures. They directly specify the lung pressure over time. When the pressure level is set to 0, no speech will be audible after executing the acoustic simulation, due to zero air flow. The f_0 tier specifies underlying f_0 targets which produce the f_0 contour. The main parameters are duration and slope, as well as starting and ending pitch (in semitones or Hertz).

The modeling of fundamental frequency is based on a target approximation model for f_0 production (Xu and Wang, 2001) and defines an underlying f_0 target for each section of an utterance. The resulting surface intonation contour is generated as a dynamic system which tries to asymptotically approximate the given targets (cf. also illustrations in Birkholz, 2007b). If a section (interval) is too short, the target is not reached but the characteristics of the curve (position, velocity) are handed over to the next section's f_0 specifications. This method is thought to represent mental f_0 targets which are not always reached in reality due to restrictions such as mass inertia.

The main strength of the gestural score is its very compact definition of numerous coordinated movements and its simplistic style which provides an efficient overview at the planning or command level.

Putting it all together: Gestural control concept

While the gestural score itself is very prominent to the user of VTL, a very important mechanism is located *behind* the actual score: the gestural control concept. This is the core component to determine the final spatial dynamics that result from the articulatory commands (gestures). It provides the link to the detailed geometric level by taking the rather abstract commands on the gestural level and translating them into complex articulatory *trajectories* and the resulting geometry of the vocal tract. Trajectories are the paths of movements in space of the individual parts of the vocal tract, or any other variables, such as f_0 contours. To work properly, the control mechanism needs data from the dominance model introduced in Section 3.1.2 (p. 31, and of course access to the vocal tract target configurations). That way, coarticulatory competition can be resolved into one sensible combined trajectory.

Its main strength is that the gestural control concept inherently includes transitions from sound to sound, which accounts for a lot of complex movement behavior in a simple way, including coarticulation, hypo-articulation and general articulatory reductions. This is, as mentioned earlier, one of the main features that makes articulatory synthesis attractive.

However, the whole control concept is a very complex mechanism, and it should be noted that it needs very meticulous configuration of the score to work properly. On the one hand, this includes gesture duration differences in the order of milliseconds, and minimal amplitude differences. On the other hand, careful attention to sensitive interactions between the tiers is needed. Most of the interactions are wanted, such as aspiration for a plosive, but some are introduced accidentally. Notably, we would like to point out that the degree of velic opening is defined in two places, firstly, in the vocal tract parameter configuration for each sound, and secondly, on the gestural tier. In general, a specification on the velic tier overrides the predefined sound configurations. From our experience, this is important to be aware of since it may lead to problems when investigating fine articulatory details and their acoustic counterparts, as will be explained in Section 3.1.6, p. 39. For a more detailed discussion of this situation and related problems, cf. Appendix A and Lasarcyk (2010).

3.1.6 Acoustic simulation

To calculate the acoustic output from the articulatory trajectories, the geometries generated by the different components described above are merged into one combined representation of the vocal tract shape. It spans everything from the lungs to the lips and nose. Since it is computationally very costly to base the acoustic simulation on complete 3D models, synthesis frameworks usually use one-dimensional simulations instead (Birkholz, 2006: Chapter 3.1). The main simplification which is made is that the movement of fluids takes place along only one dimension, i.e. on a line from the lungs to the lips, neglecting any vertical movements. The result is a representation of the vocal tract in terms of a simplified area function as shown in Figures 3.1 and 3.6.

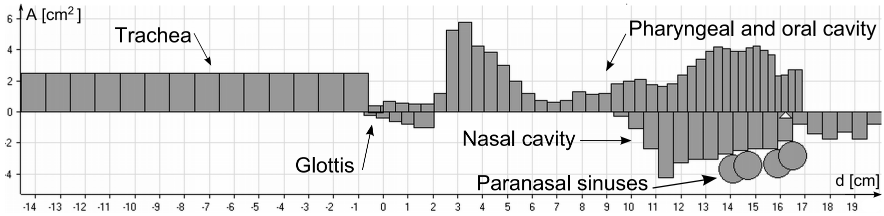


Figure 3.6: Speech tract, represented as area function (or branched tube model), from which the acoustic simulation is calculated. Annotated screenshot taken from VocalTractLab. d = Distance from the glottis along the centerline through the speech tract. Trachea, nasal and paranasal parts of the area function always stay the same. Glottal, pharyngeal and oral parts are time-varying, being based on the varying state of the glottis and the vocal tract geometry according to actual sound articulation.

This area function is the computational basis for the acoustic simulation. It is based on about 130 equidistant points along the centerline, where the vocal tract surface intersects with planes perpendicular to the centerline (for a sample area cf. Figure 3.3, left side, p. 29). The function covers the trachea, the glottis, the pharyngeal and oral cavity as well as the nasal cavity and paranasal sinuses. Taken together, these parts represent a branched tube model, which is also visualized in the software and shown in Figure 3.6. This branched tube model is the simplified structure that is assumed behind the anatomy of the speech apparatus: The air passes from the lungs through the glottis and may exit through the mouth (first branch) and the nostrils (second branch), depending on the position of the velum.

For the actual acoustic simulation, the area function (the tube model) is transformed into an electrical transmission line network on the basis of electro-acoustic analogies (cf. Kröger and Birkholz, 2007). Besides a voiced source signal the simulation network is capable of generating friction noise (turbulences) and can simulate different kinds of losses due to friction, wall vibration and sound radiation. It is thus capable of generating the sounds needed for German, if adequate vocal tract target definitions exist: Fricatives, plosives, nasals, lateral approximants, glides, and vowels. In addition to this, paralinguistic noises such as audible breathing can also be simulated. This will be exploited in the synthetic imitation of a laugh (Chapter 8).

Changes in the vocal tract geometry cause changes in the area functions and therefore lead to different acoustic sound characteristics. It is obvious that the articulatory settings in the phone configurations and their resulting trajectories constitute the primary influence on the acoustic outcome. However, there are other factors that influence the acoustic outcome. These may be less transparent but are the more important, to avoid a confounding of parameters, especially when investigating fine articulatory detail with the synthesizer.

On the one hand, these factors include detailed settings of the acoustic simulation, i.e. whether losses are calculated or neglected, various leaks are included or not etc. On the other hand, as mentioned on page 38, they include the velic settings in the phone definitions, which may be overridden by values in the gestural score. This is the case when using the synthesis mode called ‘time-domain synthesis’ (TDS). TDS is the full-fledged acoustic simulation mode that can generate all kinds of speech sounds and is controlled by the gestural score. For simple, direct vowel synthesis however, one may also employ the ‘frequency-domain synthesis’ (FDS) mode. In FDS, no individual gestural score can be defined. Instead, a predefined vocal tract configuration is used to directly create a stationary vowel sound.

When fine articulatory detail is investigated, TDS and FDS may produce critically different acoustic outcomes from the same underlying phone definition due to different values of the position of the velum: one in the original phone configuration, and one in the gestural score. To make experiments reproducible, it is therefore desirable for acoustic settings to be reported in (more) detail when describing research with articulatory synthesizers. An overview of our default synthesis profile (basic simulation settings) is provided in Appendix Section A.1.

3.2 VocalTractLab in this thesis

After having presented relevant technical components of VocalTractLab, we return to a more general perspective and sum up the appealing features of VTL for articulatory experimenting, also discussing general technical issues of VTL as regards our thesis work (Sections 3.2.1, 3.2.2), and briefly point to other work with VTL (Section 3.3).

VocalTractLab is very appealing for articulatory-phonetic research due to its high-level, articulatorily based control concept, which has inherently attractive consequences for articulation simulation. One of the most prominent aspects is the automatic generation of transition movements between targets (sounds), so that coarticulation processes do not need to be spelled out explicitly. Furthermore, since control is possible on every level, VTL offers large degrees of freedom to design an utterance, both on a global level (for pathologies or accents), or on a local one (e.g. the degree to which a target is reached during single reduced syllables). This also makes manipulations such as changing speaking rate very easy, and they come without the usual artifacts known from other synthesis methods. Even creating a singing voice is free of the usual artifacts known from corpus-based synthesis (Birkholz et al., 2007b). The synthesizer copes well with articulation at high pitch and singing-specific phenomena such as tremolo, without having to prepare specific voice databases for this task. VTL can also be used to create complex non-speech verbalizations such as laughter, which place different demands on the whole speech apparatus (cf. Chapter 8), such as rapid sequences of glottal gestures, or breathing and friction noises. This rather atypical usage of a speech synthesizer requires an increased level of fine adjustments of aspects in articulation which may not be so important in ‘regular’ speech. Lastly, the graphical output allows for easy visual inspection of articulation movements to complement findings in the auditory domain.

The question that remains is: How well does VocalTractLab function as a tool for speech articulation research, especially with regard to fine articulatory detail? This is the main *technical* evaluation question of this thesis. VTL's *general* adequacy as a research tool is based on the argument that the software brings together various established scientific models, and is based on real speaker data. Therefore its suggested articulatory output should in principle be plausible. Thus, we use VTL as an articulatory data generator, which has the advantage of being able to convert the articulatory data into corresponding speech output. This enables us to work in the auditory domain to evaluate the suggested articulatory data. Within a series of experiments we will put a number of the parameters introduced earlier in this chapter into focus for evaluation. An overview of the main technical parameters in each experiment can be found in Section 4.3, p. 66, which also provides an overview of the individual experiments.

It should be noted that the validity of VTL's internal models is a key point in the synthesis framework. Although the overall behavior of VTL suggests a general validity of the results, statements such as in Mennen et al. (2010: 17), relating to empirical contributions such as in Esling (2005), make us aware of the possibility that "our understanding of articulation and its relation to acoustics is still rudimentary, and might require wholesale revision in places." (Mennen et al., 2010: 17) Esling (2005) reports that possibly the two-dimensional, trapezoid scheme of vowel articulation needs to be revised based on findings which suggest a critical contribution of pharyngeal and laryngeal mechanisms to back vowel production.

3.2.1 Versions of VocalTractLab

As mentioned in Section 2.4, VocalTractLab is under ongoing development. Thus the software has changed since we first employed it, and we used two major versions of it in the course of this thesis. The experiments presented in chapters 10 and 11 use the version described here[2], while the remaining experiments use the previous version of VocalTractLab[3]. The overall control schemes and output characteristics of the two versions are comparable within the framework of our experiments. There are, however, some changes that will become obvious in the various figures and descriptions throughout the thesis, and they are listed here.

The vocal tract shape of the older version is not yet adapted to a specific speaker and thus has a more generic shape which is taken from descriptions in the literature (Birkholz, 2006).

The modeling of the fundamental frequency is based on a model developed by Fujisaki and Hirose (1984) in the older version. f_0 specifications are defined on two different tiers of the gestural score instead of one. The first tier controls short term variations using accent components, the second one controls long-term variations of f_0 using phrase components.

The degree of glottal opening is defined by two variables in the older version ('open' and 'close' gestures) instead of providing only the degree of glottal abduction as a target ('open'

[2] VocalTractLab 1.0, publically available at http://vocaltractlab.de/download-vocaltractlab/VocalTractLab-v1.0.zip.

[3] TractSyn, publically available at http://vocaltractlab.de/download-tractsyn/tractsyn.zip.

targets). The older version seems to allow for slightly more abrupt gestural movements at the glottis than the single-variable control interface.

Lastly, the gestural concept for the supraglottal articulation differs regarding gesture offsets and amplitudes. The older version implements the concept of gestural scores as can be found in Browman and Goldstein (1992), while in the current version the target approximation model (Xu and Wang, 2001) is implemented. In the older version, the gestures come with static targets, individual amplitudes, onsets and offsets. The offset specification elicits explicit movements towards a neutral resting position. Thus, the gesture for a sound is bidirectional: Towards a target and explicitly away from it afterwards. This is not the case in the current version of VTL, where only unidirectional gestures are used, and the lack of a vowel gesture on the score is only an implicit command to move to the neutral resting position.

Since the amplitude of a supraglottal gesture can be individually specified in the older version, one can induce hyperarticulation, which is currently not possible in the newer version of VTL. It only offers a default amplitude for supraglottal gestures. Further development is planned to again introduce a mechanism for systematic hyperarticulation within the new framework (P. Birkholz, pers. comm.).

Recently, the latest version of VocalTractLab was released (VTL 2.1, Birkholz, 2013b) which incorporates several profound changes but leaves the basic framework as it was before (cf. Section 12.3). Thus, the basic capabilities are still comparable but details within the individual modules have changed or have been enhanced. As a result, the findings from our experiments are applicable to the newest version of VTL as follows: Where VTL worked well before, that has not changed. Where VTL showed limitations in our experiments, they may have been overcome in the latest release.

More specifically, improvements can be found in the articulatory model because two anatomical parameters have been adapted to match the human physiology more closely. The mandible is now controlled only by JA (jaw angle) and HX (horizontal position), but not HY any more. The velum now has one additional parameter for better control. The course of articulatory movements can be controlled at the same level of precision as before, and the acoustic simulation has not changed – thus the basic quality of the speech output and articulatory trajectories is comparable in all versions of VTL.

3.2.2 Summary of relevant simplifications and limitations

In the following, we briefly summarize VTL's relevant limitations for this thesis. Trills can currently not be synthesized, therefore the German consonantal /r/ sound can only be synthesized in its fricative variant and not as a uvular trill.[4] The gestural scores are hand-crafted, therefore it is a time-consuming and non-trivial task to design words and longer utterances. Adequate coarticulation in consonant clusters is still hard to achieve, making words with consonant clusters less intelligible. Lastly, voice personae other than those resembling the

[4] The vocalized /r/, i.e. [ɐ], is of course not affected by this limitation

adult-male type do not sound convincing yet. These limitations influence the fundamental design of the experiments (see Section 4.1), e.g. we are not going to investigate longer utterances, and we limit the studies to utterances produced by adult males.

Apart from these apparent ‘a-priori’ limitations, we will later also discuss limitations found during the course of the experiments. It became obvious for instance that the fine-tuning of the default phone inventory regarding vocal tract and dominance variables was not yet sufficient for all sound contexts. It would cause signal distortions when attempting to synthesize particular sounds in specific phone contexts (cf. Chapter 11).

3.3 Other applications

Despite the limitations listed above, VocalTractLab currently seems to us one of the most remarkable articulatory speech synthesis systems. It provides very high-quality acoustic output, is based on a real speaker, and incorporates a three-dimensional model of the vocal tract. It is being used increasingly often as a research tool, covering a range of different topics, such as investigating consonant clusters or details of the vocal tract shape for specific single sounds, speech inversion, speech acquisition, or different processes at the glottis during singing; also, individual modules are used in isolation to study e.g. the acoustics at the vocal folds (Birkholz, 2012, pers. comm.). Other applications are presented in the remainder of this section.

With its 3D model, VTL provides the facilities to create high-quality three-dimensional graphical output which can be used for various applications besides simply providing the computational basis for the acoustic simulation: It can in principle be used to produce audio-visual speech output to create stimuli for audio-visual testing. Additionally, the 3D component can be exported into separate programs for developing e.g. virtual talking characters (avatars), whose lips, tongue and larynx movements then look more naturalistic when moving correctly according to the speech sounds uttered (see e.g. Walter, 2006).

The visual component can furthermore be used for speech therapy and teaching applications. Visualizing the movements within a transparent vocal tract at any desired speed, i.e. also in slow motion, can help students or patients with speech sound disorders to understand the dynamics of speech articulation. Especially for phonetic and logopedic education, the different visualization features of VocalTractLab are a useful supporting medium when learning about relations between articulation and acoustics. For instance, one can directly visualize the influence of voice quality on the spectrum; or one can demonstrate the difference between formants and harmonics by separating the influences of supraglottal articulation vs. voiced source signal manipulations.

Although so far we have only introduced the gestural *score* as a means to control articulation, the movements can also be controlled more directly, i.e. at a lower level, namely by articulatory data. The synthesizer then imitates predefined movements directly, without executing any gestural control concept calculations. When the descriptions of the movements are provided in terms of electro-magnetic articulography (EMA) data, we speak of resynthesis from EMA data (Birkholz et al., 2007b, Steiner, 2010). In this technique, it is the EMA data that control the deformations of the vocal tract, and not the gestural score. VTL can then serve as a visualization-evaluation tool for management of large amounts of data. It visualizes the articulatory movements within its model articulators, and additionally, it offers acoustic simulation to evaluate the complex nature of the EMA data in the auditory domain: When the sound characteristics are close to an expected speech pattern, it indicates appropriate structures in the EMA data. Ultimately, gestural patterns could be extracted from appropriate data to automatically learn articulatory movements that correspond to specific phones and phone sequences (see also Preuß et al., 2013).

Because of its high-quality aerodynamic-acoustic simulation models, the development of VTL represents a step towards a potentially broader (and perhaps in the long-term commercial) use of the articulatory speech synthesis method. It seems that, besides important 'local' enhancements which are subject of ongoing research,[5] the one major step towards numerous applications would be to provide full text-to-speech (TTS) capability. With an automated, complete TTS synthesis process, VocalTractLab would be capable of synthesizing a broad range of utterances in a fast and easy manner. This will of course only work if the segmental quality is reliably good. To date, the friction noises in particular still need improving. Therefore, it can only be a mid- to long-term goal to implement a full TTS system.

We believe that having a TTS component would be an important milestone. As soon as it meets a certain level of intelligibility it can be broadly used, conveying the desired linguistic messages. Complementary to that, it is nevertheless important in the long run to be able to control and to convey paralinguistic features. Only this can lead to a more naturalistic sounding synthetic voice. Therefore, in this thesis, we focus on articulatory-phonetic details of paralinguistic phenomena to better understand them and to be able to replicate them.

[5] These include improved burst explosion modeling, improved consonant cluster coarticulation, a new coarticulation model, and easy-to-use voice quality settings (Birkholz, pers. comm.). See also Section 12.3.

Chapter 4

Aims and methodology of the experiments

After having laid out the background on relevant aspects of speech production and speech synthesis, we turn to the empirical part of the thesis in this chapter. We introduce the main assumptions and procedural aims (Section 4.1) which led to the common design of the experiments (Section 4.2), and conclude with a brief summary of the topics of each experiment (Section 4.3).

As stated in Chapter 1, the overall goal of the thesis is to investigate paralinguistically motivated fine articulatory adjustments, and, related to that, to assess the technical capabilities of the synthesizer that enable the simulation of these fine articulatory details. We selected a close-copy by imitation method for carrying out these investigations, with the goal of obtaining synthetic speech samples that are related as closely as possible to the acoustics of real human speech data. The rationale behind this is that when we achieve a close acoustic imitation by using an articulatory synthesizer, we simultaneously obtain articulatory data for free along with it. This has been laid out in Section 2.5, explaining how VTL can be used as a data acquisition tool. The system-inherent articulatory transparency can be used to investigate articulatory processes within various articulatory-phonetic research questions.

4.1 Assumptions and procedural aims

The main assumptions that underlie the methodology used in this thesis are the following:
Assumption 1: As stated in the thesis Introduction, our basic perspective is that the synthesizer brings together and organizes known empirical articulatory knowledge. It is capable of producing intelligible speech by controlling parameters that are transparent, intuitive and articulatorily grounded. We take this as an indication that the basic requirements for a functional articulatory synthesis system are met. These requirements are that scientific theories and models have been adequately implemented in the synthesizer. For our procedure this means that in general we can trust the articulatory suggestions contained in the output of VocalTractLab.

Assumption 2: We assume further that, from the documentation of the synthesizer, we are sufficiently aware of the model's simplifications as sketched in Section 3.1 (e.g. that it incorporates a geometric and not a physiological model of the vocal tract). Therefore we can project eventual consequences from these simplifications onto our speech samples. This means that for certain phenomena, we can relate our results to these simplifications because we know of the current limitations of the synthesizer. For instance, awareness of these simplifications advises us to explicitly check each final imitation for articulatory plausibility by means of visual inspection (e.g. vocal tract shapes of static target sounds).

In the following we introduce and discuss the procedural goals of this thesis, including some consequences arising from them.

Goal 1a: We take human speech data as a role model, and wherever possible we investigate properties of real-life spontaneous utterances that are available in existing speech corpora. If this is not feasible, we use speech that is recorded for the purposes of the experiment.

Goal 1b: We are committed to the primacy of human articulation, that is we want to investigate articulatory processes that correspond to human speech articulation or that are plausible as human articulation. In other words, we do not want to produce articulatory data for a possible way of speaking which may resemble alien articulation just because the synthesizer is capable of it. It might sound the same but it is not relevant to the investigation of human articulation.

Since VTL is currently not capable of simulating certain sounds, such as vibrants (as summarized in Section 3.1, p. 42), we are not going to investigate these sounds: We would perhaps be able to come up with an acoustic imitation but it would not represent the usual human articulation pattern. It would rather be a technical preliminary solution with matching acoustic properties. Similarly, we use, but do not investigate deeply, sounds that are known to be difficult for the current implementation of VTL. These include e.g. plosives due to the current models used for burst simulation. It is not feasible to avoid these sounds completely because it would restrict the set of possible wordings for utterances too much. Lastly, we do not deal with female or non-adult voices since the synthesizer is currently offering a configuration only based on an adult male speaker.

Goal 2a: We place the judgment of the naive listener into the focus of our evaluation. Although we evaluate *articulatory* speech synthesis, we believe that in the end it is the perceived speech wave that should be the primary domain of evaluation regarding the synthesis system. Thus, approval of an imitation in the perceptual domain is used as the prerequisite for carrying out further analyses with respect to suggested articulatory movements. That is, when an imitation sounds good (and is generally plausible in its basic articulatory settings), we analyze its fine articulatory details with respect to the given research question. If it does not sound good, it has to be optimized or the reasons why optimization is not successful have to be explored.

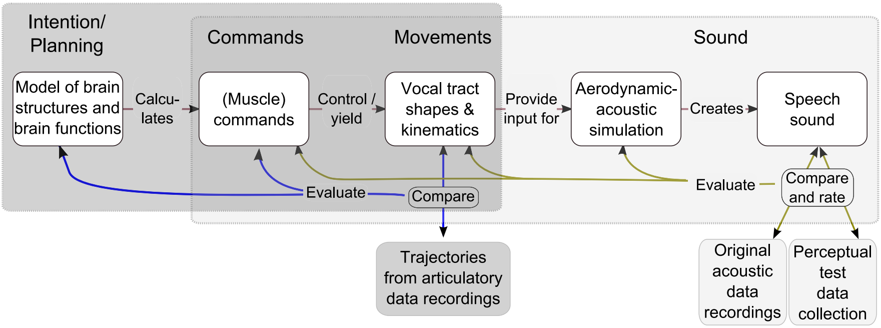

Intention/ Planning
Commands
Movements
Sound
Model of brain structures and brain functions
Calcu-lates
(Muscle) commands
Control / yield
Vocal tract shapes & kinematics
Provide input for
Aerodynamic-acoustic simulation
Creates
Speech sound
Evaluate
Compare
Evaluate
Compare and rate
Trajectories from articulatory data recordings
Original acoustic data recordings
Perceptual test data collection

Figure 4.1: Illustration of two different methods in speech production research, comparing two evaluation cycles (dark vs. light gray). One type of speech production research (dark gray) is carried out in the domain of speech planning via the issuing of commands for the control of the articulators. The models are often evaluated against natural articulatory data. The other type (light gray) – applied in this thesis – starts on a command level and evaluates synthesizer models in the acoustic domain: The main criterion for evaluation is the auditory acceptability of the synthesized speech. It can be seen as a complementary evaluation strategy, building on the original knowledge from articulatory data studies and filling in more articulatory details by letting the synthesis model generate possible trajectories and vocal tract shapes.

Goal 2b: We use acoustic data as the primary resource for imitation. Therefore, we use acoustic recordings of speech which we then imitate, and optimize the imitation's details until close acoustic resemblance between original and synthesized version is reached.

Grounded in Goal 2a and 2b, the final evaluation of the articulatory suggestions made by VTL is based on perception tests with naive listeners, and the early drafts of the imitations have to pass informal immediate evaluations with the aim that what does not sound good must be reworked. In other words, the evaluation is not based on machine metrics such as signal comparisons by means of high-dimensional vectors, and is not based on evaluating alignment to natural articulatory data. In some cases, selective acoustic analysis seems reasonable in order to be able to provide further explanations e.g. regarding (seemingly) opaque perception test results.

This acoustic-perceptual approach can be regarded as a complementary approach to research methods based more on articulatory data (cf. Figure 4.1). We conceive these methods as being interested more strongly in the higher levels of speech production such as speech planning mechanisms in the brain, while for our research, the planning stage is represented as a technical module in the form of gestural scores. In the former methods, the evaluation of hypotheses is typically carried out by comparing the model articulatory trajectories with human articulatory data, whereas we pursue techniques of evaluation that use the *acoustic* domain for similarity or acceptability assessments.

We assume that, for different reasons, neither approach can be 'complete' but each of the research methods has specific aims for which the corresponding approaches and evaluation methods seem adequate. On the one hand, the speech production research concerning the higher levels of speech production, such as speech planning, focuses on brain activity and its effects in terms of nerve impulses – or abstractions thereof – on tissue and muscles, in order, ultimately, to improve the modeling of these parts of the human speaking system. Usually the articulatory databases (nowadays) only show the trajectories of one or a limited number of speakers and a limited number of utterances. Therefore, they document a subset of all possible articulatory behaviors, and depending on the data it is difficult to infer how usual or unusual the recorded movements are compared to the general population of a speech community.

On the other hand, the approach we use here is focused on communicative efficiency – when basic articulatory plausibility is attested – and therefore is interested in the perceptual effects of certain articulatory behaviors, leaving aside any modeling of intentions and preparation of speaking in the brain. An imitation must sound good and be intelligible, thus the auditory system of naive listeners is the evaluator for our articulatory suggestions. However, when working mainly in the acoustic domain, we have to keep in mind that any one acoustic outcome can be obtained by different articulatory behaviors (many-to-one mapping problem, cf. Section 2.2), and as with the other method, we often have no direct way to assess how usual or unusual our articulatory suggestions are for a given speech phenomenon.

So both frameworks provide insights into speech production based on single examples, i.e. the reference speech data, be it articulatory or acoustic. However, both frameworks are open to more extensive data processing, so that more general statements about different aspects of speech production could be obtained in the future.

4.2 Procedure within each experiment

The central goal of the experiments is to investigate a series of phonetic research questions, to gain insights into articulatory processes during speech production. Besides this, the experiments also have a technical evaluation goal, to give an indication of how well different articulatory processes are simulated in the articulatory synthesis system.

Each of the experiments follows the same experimental method, which could be described as ‘articulatory imitation and evaluation’. The main steps of this method are – after some preparatory steps – articulatory imitation and plausibility assessment, combined with formal auditory evaluation and acoustic characterization. We first give an introductory overview of this process. The following subsections then explicate details on the particular workflow used at the different stages of each experiment. A visual overview of the workflow is presented in Figure 4.2.

With respect to a particular phonetic research question, we study a given phonetic target phenomenon from a speech database by analyzing the sound signal and gathering phonetic (acoustic and/or articulatory) descriptions in the literature to formulate hypotheses about the articulation of the given phenomena. These are transferred into the speech synthesizer’s

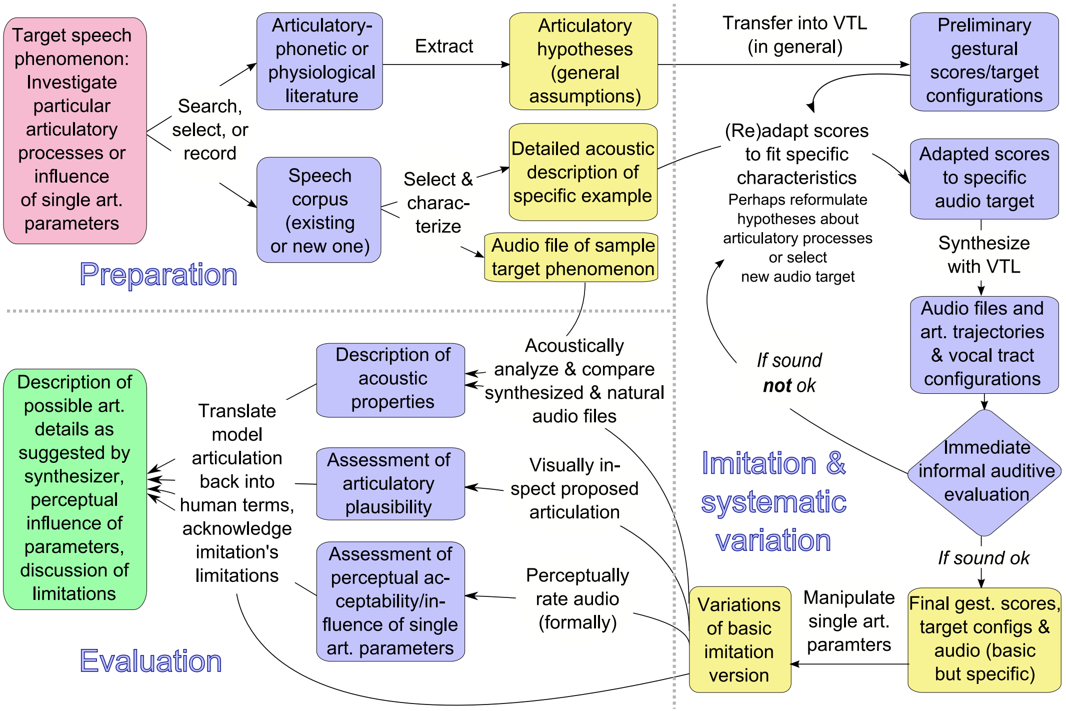

Target speech phenomenon: Investigate particular articulatory processes or influence of single art. parameters
Search, select, or record
Articulatory-phonetic or physiological literature
Extract
Articulatory hypotheses (general assumptions)
Speech corpus (existing or new one)
Select & charac-terize
Detailed acoustic description of specific example
Audio file of sample target phenomenon
Preparation
Transfer into VTL (in general)
Preliminary gestural scores/target configurations
(Re)adapt scores to fit specific characteristics
Perhaps reformulate hypotheses about articulatory processes or select new audio target
Adapted scores to specific audio target
Synthesize with VTL
Audio files and art. trajectories & vocal tract configurations
Immediate informal auditive evaluation
If sound not ok
If sound ok
Final gest. scores, target configs & audio (basic but specific)
Manipulate single art. paramters
Variations of basic imitation version
Imitation & systematic variation
Acoustically analyze & compare synthesized & natural audio files
Description of acoustic properties
Visually in-spect proposed articulation
Assessment of articulatory plausibility
Perceptually rate audio (formally)
Assessment of perceptual ac-ceptability/in-fluence of single art. parameters
Translate model articulation back into human terms, acknowledge imitation's limitations
Description of possible art. details as suggested by synthesizer, perceptual influence of parameters, discussion of limitations
Evaluation

Figure 4.2: Basic structure of every experiment: A specific speech phenomenon is analyzed on various levels (acoustic/articulatory, recorded data/findings in the literature) and used to build hypotheses about its articulation. These are then implemented, so that the articulatory synthesizer creates an imitation of the target sample, which can be modified in individual parameters to generate different versions of the original phenomenon according to the research question. The resulting acoustic stimuli are analyzed for acoustic/articulatory adequacy, and for perceptual acceptance or salience of individual articulatory parameters, thereby evaluating the articulatory hypotheses of the target speech phenomenon and also synthesizer itself (cf. Sec. 4.2). *art.* = articulatory, *gest.* = gestural.

control parameters (gestural scores, vocal tract target definitions), thereby connecting the acoustic and the articulatory domain. The acoustic stimuli that are created from these hypotheses represent basic synthetic versions of the initial phenomenon and suggest articulatory movements associated with the imitated speech. From these basic versions we can create systematic articulatory variants to test the effect of different parameters. The output is, firstly, assessed regarding articulatory plausibility, secondly, analyzed acoustically and, thirdly, tested perceptually whether the salient target features are conveyed to a naive listener. The individual evaluation aspects are then combined into a final conclusion, firstly about the phonetic research question(s) and secondly, about the performance of the articulatory synthesizer, to answer the technical research question.

The design follows a paradigm of imitation and evaluation, similar to two other paradigms found in speech synthesis research: Copy synthesis and analysis-by-synthesis. Copy synthesis is a method of analyzing an acoustic speech signal and extracting relevant information to create an imitation of the sample by resynthesizing it. This usually takes places in the acoustic domain, meaning that an acoustic representation of the sample utterance (i.e. the speech signal) exists and we create as close a synthetic version of it as possible (cf. e.g. Scheffers and Simpson, 1995, Laprie and Bonneau, 2007). Copy synthesis can be contrasted to text-to-speech synthesis. TTS synthesis creates an utterance from a given text. The acoustics of course have to be appropriate (to be intelligible) but the focus usually lies on the linguistic message that one wants to convey.

This technique can be combined with or is subsumed under the analysis-by-synthesis approach, where a natural speech sample is selected to be imitated by means of a synthesizer (most commonly formant synthesis, such as with systems based on Klatt, 1980). Then one can analyze which parameter combinations shows the best results or the closest matching in terms of visual comparisons of spectrograms, acoustic measures, or perceptual evaluations. Additionally, since one can control signal parameters individually, one can test their influence on the acoustics and perception independently. Examples are numerous and can e.g. be found in Gupta and Schroeter (1993), Gabelman et al. (1998), Kreiman and Gerratt (1996), Antonanzas-Barroso et al. (2005) etc.

The methodology we use in the experiments is based upon these approaches and adds the articulatory domain to them. Instead of tweaking acoustic parameters on the signal surface as in traditional copy-synthesis, we adjust gestural commands which in turn change the acoustic results. The rationale of the analysis-by-synthesis paradigm is likewise transferred into the articulatory domain: Articulatory parameter combinations, and not acoustic ones, are tested for their perceptual effects.

For the understanding of the process of imitation, it is important to note that we generate three kinds of imitations which build up upon each other. The first one is a *basic, general imitation* of a natural speech sample which implements the general articulatory hypotheses gathered in a preparatory step. The fine temporal and intonational structure which is present in the natural sample is imposed onto this general imitation to create the *adapted, detailed* *imitation*. Most of the experiments involve a third step, in which the adapted imitation is systematically manipulated in certain articulatory parameters to create *systematically varied imitations* which serve as test items in the perception tests.

An advantage of the overall empirical paradigm lies in the fact that it is based on relatively easy-to-acquire data produced by simulation. In using the articulatory synthesis software as a *bridge* between acoustics and articulation, we can set up a framework for the experiments that allows for possible articulatory insights without having to resort to the much more costly articulatory data acquisition methods (cf. Section 2.1.3). The software itself is based on articulatory and anatomical data combined with acoustic models, and therefore can be regarded as a scientifically based tool to explore speech articulation and investigate hypotheses about articulation details. The hypotheses can be tested by isolated manipulation of parameters, which is only possible in a virtual vocal tract, not in a human one, even if the speaker is a trained professional. By conducting listening tests, the architecture of the experiments allows both for evaluation of individual articulatory parameters and for evaluation of overall acoustic stimulus quality. At the same time, it enables us to evaluate the synthesizer itself by reporting technical capabilities or limitations that become apparent during the process of stimulus creation.

In the following, we describe the individual steps of the workflow in more detail.

4.2.1 Knowledge base and selection of audio target

The input to an experiment is a research question and an idea of an appropriate (representative) target phenomenon. The research question determines which desired articulatory variations or basic articulatory processes are going to be the focus of the investigation.

Before the actual imitation and evaluation can start, we need a preparatory step of gathering knowledge and audio material. In each empirical chapter, we introduce the phonetic topic by giving a brief overview of the articulatory hypotheses found in the phonetic-articulatory or physiological literature according to the focus set by the phonetic research question. After extracting and organizing known relevant articulatory details from the literature, we develop general assumptions about the basic articulatory processes taking place. These will later guide the creation of the *general* imitation of the natural speech sample.

At this point, it may become apparent that some aspects of articulation have not yet been investigated in sufficient detail, or that descriptions are not available to us. So the question is whether the knowledge suffices for resynthesis. In these situations, we make articulatory assumptions from context and suggest an ad-hoc working solution to obtain the desired speech phenomenon (as in the case of the basic articulatory processes during laughter, cf. Chapter 8, p. 117). The evaluations at the end of the experiment provide a first assessment of the adequacy of the ‘stop-gap’ assumptions.

Since it is our aim to investigate real speech phenomena, we use audio recordings from natural speakers as a target utterance and copy-imitate them as closely as possible (Goal 1a).

If possible, we try to find speech samples from spontaneous speech. For some topics, such as laughter and laughed speech, good examples can probably *only* be found in spontaneous speech (if one excludes acted vocalizations). If suitable speech corpora are available, we search and select single audio utterances that fit the target phenomenon best. Selection criteria are:

- The subjective auditory impression: Which sound samples do best represent the target phenomenon?
- A rough assessment of articulatory complexity: Which of these sound samples are articulatorily 'simple' enough to be imitated by VTL? As stated in Section 3.1, we e.g. do not concentrate on speech sounds that are currently not fully available in the synthesizer such as uvular vibrants. From early synthesis trials it also became clear that complex consonant clusters are very hard to synthesize accurately and therefore we try to avoid target samples with a segmental structure that is too complex.

For some experiments, we designed and recorded our own speech corpora and selected the target audios in the same way as described for the existing speech databases.

After having selected a natural speech sample for imitation (the 'role model'), we characterize the audio signal acoustically in a very detailed way, i.e. listing f_0 contours, formant structures, intensity levels, segment durations, etc. This description serves as a guide to the *detailed*, adapted imitation of the later stimuli.

It often happens that a speech sample that initially sounded appropriate proves to be unsuitable for imitation, but no satisfactory alternative is available. Suitability is affected when noise or other voices are audible on the audio track because, typically, clean natural audio samples are needed for baseline comparisons in the listening test evaluation or for reliable formant analyses. So in these case of data sparsity in the corpora, we e.g. redesign the experiment, look for similar other target samples, or record our own audio corpus.

At the end of this preparatory step, we present hypotheses about general articulatory processes or presumed vocal tract target configurations and an illustration of the possible articulatory process as a whole, to guide the *general* imitation of the target speech phenomenon. Additionally, we select a target audio recording (wave file) and present a detailed description on the acoustic level, which guides the *adapted* imitation. The audio file also serves as a reference during potential comparative acoustic analyses, and as a potential baseline stimulus during formal listening tests.

4.2.2 General imitation

In this step, we produce a preliminary imitation of the target phenomenon. It may be regarded as the central step of the whole method because it involves setting up the first comprehensive alignment of articulatory actions within the synthesis model. The input to this step are the general, rough hypotheses about articulation stemming from articulation analysis as described above. These articulation assumptions are transferred into the format of the

VTL gestural scores. In some cases, the aim is to create new vocal tract target configurations from existing similar ones, so besides the gestural scores we develop new vocal tract target configurations with the desired articulatory or acoustic properties. When a preliminary basic gestural score and perhaps new vocal tract target configurations have been established, we run the synthesis process (aerodynamic-acoustic simulation) to obtain the corresponding audio files and articulatory trajectories.

Generating a basic imitation of the natural speech sample always includes an immediate informal auditory evaluation by two or more trained phoneticians (Goal 2a). This includes assessing the overall quality of the imitation, its similarity to the natural speech sample, and checking for audible signal distortions which might indicate inappropriate articulatory movements such as too abrupt changes of articulators. In some cases (smiled speech and regional accents) we also asked a number of speech science experts to produce narrow transcriptions of the vowel qualities they perceived to document the auditory-acoustic quality of the sounds.

If simulation results are auditorily inappropriate or not as expected, iterative fine-tuning to the gestural scores is applied, and the synthesis is re-run, until an appropriate sound is obtained. In most cases, an imitation goes through many iterations until the final basic version is obtained. To give an impression of this gradual development in some experiments (e.g. during the imitation of words in Chapter 11), we give short characterizations of the current acoustic state of an imitation to make obvious the wide spectrum of (‘inappropriate’) segmental, but also prosodic quality, and describe paths taken until the final gestural scores have become established.

The transfer of the general articulatory hypotheses into specific articulatory commands on the gestural score is not always a straightforward process. It is exactly at this point that it becomes obvious if some desired gestural patterns are hard to implement. This might partly be the case because the articulatory descriptions are not couched in terms which conform directly with the synthesizer’s control parameters. But in general, the parameters are intuitive enough to successfully transfer articulatory plans into the system.

This step is also the primary location at which potential mismatches or imprecisions (gaps in the articulatory descriptions) become obvious in the articulatory hypotheses or the technical ‘translation’. As a result, the initially proposed gestural score fails to produce appropriate acoustic results. This leads to changes of plans, which can be three-fold: The smallest changes include adjusting the gestures or vocal tract target configurations in relevant details. More deep-seated changes mean reformulating the articulatory hypotheses altogether, and if this is not successful, the third option regarding change in plans is to select a new target audio sample, e.g. one with a less complex segmental structure if segmental pronunciation is the problem, and rerun the imitation process with a different audio role model.

The methods for building the basic versions of the gestural scores are mostly manual and iterative in nature. We build the gestural score from scratch, guided by the above-mentioned articulatory hypotheses, and transfer them to the synthesizer’s gestural format. In some cases, we use a semi-manual method, by applying a singing-TTS function of VTL, which is described in Birkholz (2007a). It is an implementation of a set of rules that generate the basic articulatory patterns on the gestural score for each syllable specified. A syllable is ‘sung’ on a specified note (pitch) for a given duration (note length) with given lyrics (one or more speech sounds given in SAMPA notation). The result has to be manually adjusted in every case, firstly because of differences in intonation (singing vs. speaking) and secondly because of remaining distortions in the segmental quality. It is nevertheless a very useful starting point for score development because the basic skeleton of the gestures does not have to be created manually.

In summary, this step results in a *general* imitation of our target speech phenomenon. Acquired material are the gestural scores and vocal tract target configurations as well as corresponding audio files and articulatory trajectories. The general benefit from this is that we have acoustics and articulation that correspond to each other, therefore providing one possible complete[1] transparent mapping of acoustics to articulation and vice versa.

4.2.3 Specific (adapted, detailed) imitation

The general imitation now needs to be adapted to fit the acoustics of the natural speech sample because our goal is to imitate real speech. Additionally, an adapted imitation can readily be used side by side with its role model in baseline-vs.-test-condition experiments. To this end, the specific acoustic properties of the original audio signal, which have been analyzed and described before, are integrated into the gestural scores or vocal tract target configurations. This means adapting the existing general (generic) gestural score so that it matches desired segment durations, f_0 contours, relative intensity levels etc., or adapting the default sound configurations to match desired spectral qualities, such as regionally accented vowel qualities.

From this step, we obtain imitations that are adapted to a specific acoustic target. Again the available material includes gestural scores and target configurations as well as corresponding sound files and articulatory trajectories. Thus, now we have articulatory transparency for the specific natural speech sample which represents the target speech phenomenon that is part of the initial research question for an experiment. The importance of this step is that by running the synthesis algorithms, the intuitive (qualitative) knowledge about articulation is complemented by a detailed (quantitative) provision of articulatory data. Of course, this is based merely on the models that are implemented in the synthesizer. However, since we presume that the models are implemented correctly, this step nevertheless means that we are able to close – or at least narrow considerably – the gaps in the initial articulatory descriptions and hypotheses, the answers being based on a specific articulatory modeling as used in the VTL software framework (Assumption 1, p. 45).

[1] Complete in the sense of that the whole vocal tract is described and not only parts of it.

4.2.4 Systematic variants

Depending on the research questions, most experiments involve testing the influence of a particular articulatory parameter on listeners' perception. For this, we adapt the relevant parameters within the possible and plausible ranges of the articulatory model, e.g. manipulate the degree of lip spreading to investigate characteristics of smiled speech. Our manipulations can take place in two locations:

- The gestural scores (gestural sequence of events)
- Their associated vocal tract targets (underlying phone definitions)

For each desired parameter combination, one separate sound file is created from our *adapted* version of the imitated speech phenomenon by manipulating individual parameters. These sound files then represent systematic articulatory variants of the adapted imitation. In most cases, the systematic variation takes place in the phone definitions, in some cases (such as with the laugh experiment) we manipulate the number and placement of gestures on the gestural score. Score manipulation is always done manually. In the following, we discuss some points regarding the manipulation of the phone definitions.

The methods for creating new vocal tract target configurations (the macros to define a phone) are mostly manual in nature, similar to the generation of the basic gestural scores described earlier. This means that we manipulate the relevant vocal tract parameters in the phone dialog (GUI) or the configuration file according to the articulatory properties that we wish to simulate (such as lip spreading). We thereby create systematically varying versions of a predefined phone. For instance, to create a series of 'smiled' vowels, we increase e.g. lip spreading or larynx height to various degrees in different combinations. Each combination is saved as one distinct new sound configuration which can then be integrated into the gestural score by using the appropriate labels in the score's vowel gestures.

Batch processing

Prior to creating systematic variants, it is often necessary to determine plausible parameter ranges or adequate typical combinations of parameters. Finding these configurations is mostly done manually by setting parameter values and listening to the results. In one experiment however (vocal aging, Chapter 7), we were able to use a batch tool contributed by Michael Feld from the German Research Center for Artificial Intelligence DFKI (Feld, 2011). It accelerated the process of stimulus creation considerably because we were able to create and listen to a broad range of systematically differing stimuli much more quickly, in order to decide which configurations should be used for formal evaluation.

One important aspect to note here is that creating a new phone includes defining appropriate dominance values for each parameter in the phone definition as well. In general, the creation of new vowels is trivial as far as the dominance values are concerned, since all values are set to 100 percent. For consonants a more individual dominance profile is required.

Some of the dominance values are extremely sensitive and require a lot of adjustment. This becomes obvious when the new phones are integrated into words, since the dominance values define which phone may occupy which articulator most strongly in a competitive situation. We document more details of this fine-tuning in Chapter 11, where the creation of regionally accented words reveals the need for adjustments in both vowel and consonant macros.

Constrained acoustic-to-articulatory inversion

Creating new target configurations, i.e. new phones that fit specific characteristics can also be done with an automatic algorithm. One example is the *formant optimization algorithm* of VTL, which was developed further by Peter Birkholz for application in this thesis. It calculates several possible vocal tract shapes for a given triple of formant frequency values that serve as an acoustic target (F_1, F_2, F_3).

This algorithm represents a method of acoustic-to-articulatory inversion. Since inversion faces the *many-to-one problem* (cf. Section 2.2), one acoustic output can stem from different articulatory shapes, i.e. the vocal tract shape that produces given formant frequencies is not unique. Therefore all articulatory solutions of the algorithm, i.e. the proposed vocal tract shapes, can only be seen as plausible sample solutions. During the calculations, the parameter space is searched to find *a* good approximation. Since the vocal tract is described by a large number of vocal tract parameters, it is not feasible to search the entire parameter space. Instead, the well-known optimization technique of “simulated annealing” is used (Jacobs et al., 1982; Černý, 1985; Kirkpatrick et al., 1983; Kalos and Whitlock, 2008). During a fixed number of iterations, it changes the parameters of the vocal tract, starting from a given initial configuration. To find approximations for a particular vowel, we use the corresponding default vowel of the VTL-speaker configuration as initial configuration.

Since the starting configuration is relatively close to what we are looking for as a target, every run of this algorithm produces a good approximation.[2] A statistical component causes the result to vary for each run. The results represent very good *local* solutions but not a global, optimal solution. In a way this is similar to the situation in human vocal tracts. Speakers aim to produce a particular sound but different speakers do it slightly differently, and every production within an individual speaker is different too.

The optimization process in general works as defined in the pseudo-code listing below. The current status of the vocal tract parameters is stored in a parameter vector. In each iteration, the single parameters are changed by a certain small amount. The resulting vocal tract shape is used to calculate the new formant frequencies, which are then used to calculate an error value. If the error decreases compared to the old parameter configuration, it represents a better solution than the old vocal tract shape. This solution is accepted as a new starting point for the next iteration. A stochastic component is introduced to allow other configura-

[2] Since the starting point of the algorithm is a similar sounding, existing vowel definition of the standard phone set of VTL, this technique can be regarded as a *restricted-scenario inversion algorithm* since it primarily searches within the spatial vicinity of the initial vocal tract configuration.

tions to be accepted under certain circumstances as well. This enhances the space of possible solutions that can be found because the search does not operate globally. So by temporarily accepting a ‘worse’ vocal tract shape, this might build a bridge to another local space for other ‘better’ configurations.[3]

```
$T := 1.0$
$E :=$ Formant error for initial parameter vector $v$
Loop 100 times
  Change each parameter $v_i$ by a small amount to $v'_i$
  $v'_i := v_i + s \cdot \text{rand}[-1,+1] \cdot (max_i - min_i)$
  $E' :=$ Formant error for parameter vector $v'$
  $p := e^{-(E'-E)/T}$
  if $(E' < E)$ or $(\text{rand}[0,1] < p)$ then
    $v := v'$ and $E := E'$
  $T := T \cdot 0.95$
End of loop
```

T is the temperature that is decreased in each run by the factor 0.95; rand[a,b] denotes an equally distributed random number between a and b; max_i and min_i denote the maximum and minimum of the range of parameter i; p is the probability that a new vocal tract configuration with a higher error is accepted. The formant error is defined as

$$E = \sqrt{\frac{1}{3}\left[(1-\frac{F_1}{F'_1})^2 + (1-\frac{F_2}{F'_2})^2 + (1-\frac{F_3}{F'_3})^2\right]} \cdot 100\% \tag{4.1}$$

where F_1, F_2, and F_3 are the current formant frequencies and F'_1, F'_2, and F'_3 are the target frequencies for the optimization. s defines the step size for the simulation. It works best with values between 1.0 and 2.0, since prior testing revealed that these step sizes would most likely produce output with low acoustic divergence from the target values.

The domain of evaluation, i.e. how well the inversion process worked, is the acoustic domain, more precisely the formant frequencies of a given vowel. The formant error function uses the unit Hertz for the calculation of errors between old and new formants. This does not fully account for the different sensibility to frequency ranges of the human ear. By using percentages to express the error, the auditory impact is modeled at least in a simplified way. So a 6% error at 1000 Hz means the value is 60 Hz off. A 6% error at 200 Hz means the proposed solution is 12 Hz off.

The use of both manual and automatic methods to define new phones is linked to the different research questions behind the experiments. In most experiments, we have an idea of the articulatory processes or manipulations that we want to test and can therefore input the articulatory parameter values directly, testing the effects of the articulation differences in subsequent perception tests. In the regional accent experiment (Chapter 10), our aim is to imitate the typical *sound* of an accent and from this perspective examine the proposed possible corresponding articulatory adjustments. So while most of the experiments have

[3] Pseudo-code listing courtesy of Peter Birkholz.

an articulatory-phonetic goal – such as ‘additional lip spreading’ – the regional accent experiment is set up with an acoustic-phonetic goal of imitation to find out which systematic articulatory variations may take place in (this type of) accented speech.

Of course, the accent study also starts with an articulatory hypothesis; so why not proceed by manually defining new phones as well, guided by the articulatory assumptions? This leads to a second reason for employing both manual and automatic methods: The difficulty of precise acoustic target matching. In pilot studies we tried to define new phones manually, but the large number of parameters (i.e. degrees of freedom) in the vocal tract, combined with the quantal nature of speech made it unfeasible to obtain acceptable sounding vowels by adjusting the articulators to theoretically motivated positions. The algorithm, however, is designed to find these fine-grained spatial differences and therefore it is more suited than manual adjustment methods.

At this point, we have synthesized all the speech material that we need and proceed to the evaluation block, which covers three kinds of evaluation domains: acoustic, articulatory and perceptual.

4.2.5 Acoustic evaluation of the stimuli

The first kind of evaluation takes place in the instrumental-acoustic domain. We use the audio files synthesized from VTL and, if applicable, also corresponding audio files from a human speaker, and perhaps some background information about acoustic characteristics human voices usually show for given aspects of the voice. This enables us to assess whether our natural speech samples are representative enough to serve as appropriate baselines for comparison.

We now acoustically analyze the audios by retrieving different acoustic measurements with Praat (Boersma, 2001) or Wavesurfer (Sjölander and Beskow, 2000). The measurements depend on what aspects of an audio stimulus have to be characterized, and may e.g. include voice quality measurements or formant frequency values. Lastly, we compare the artificial acoustic profile with the human one (measured or expected). They should be rather similar or the divergences should be explainable, otherwise, more investigation is needed to find the source for the differences. Please note that these measurements only serve evaluation purposes and are thus different from the detailed acoustic descriptions mentioned above (f_0 contours, sound pressure levels, segment durations) which are used to guide the imitation in the first place.

In the end, we present a description of acoustic profiles and give an interpretation of these measurements.

4.2.6 Articulatory plausibility assessment

The second kind of evaluation takes place in the articulatory domain. As input we use the gestural trajectories and phone target configurations obtained from the imitation step. We visually inspect the trajectories with respect to the question whether they can possibly be executed by humans and are continuous in the time domain, i.e. do not show any ‘jumps’ between articulatory positions or otherwise manifest gestural patterns which diverge from what has been observed in articulatory studies of human speech production. Regarding the vocal tract target configurations, we visually inspect the overall vocal tract shapes to see whether they conform to plausible contours with no conspicuous deformations (Goal 1b, p. 46).

In the end we obtain an assessment of articulatory plausibility for the given imitation.

4.2.7 Perceptual evaluation of the stimuli

The third and last kind of evaluation concerns the auditory, or perceptual domain. It may be regarded as the central part of the evaluation: According to Goal 2a (p. 46), we put the judgment of the naive listener into the center of our evaluations since we believe that ultimately, the most important thing in communication is that an utterance sounds ‘good’ (adequate, intelligible, etc.). So for practically all experiments we carry out a formal perception test.

General remarks

The formal perceptual evaluation is carried out in addition to the above-mentioned immediate informal evaluation. The informal evaluation, on the one hand, is done by phonetically trained listeners by means of analytic listening and it serves as a basic overall quality monitoring to timely correct unwanted distortions in the synthesized sample or to adjust other properties of the utterance. Formal evaluation, on the other hand, gives indications about the perceptual effects of the systematic fine-articulatory manipulations. It is done by listeners that are naive to the central ideas of the research question and are often not even informed about the artificial nature of the voice.

All the perceptual evaluations demand roughly the same requirements from the participants. Most importantly, these include that the listeners are German native speakers and that they complete the whole test, or in other words, we include only complete data sets in the statistical evaluation. For most of the evaluations, we use test setups that are internet-based (see p. 60), which then require the participants to be able to use a computer with audio output and to have access to a (high-speed) internet connection.

The audio material that we use are the audio files obtained from the imitation steps, e.g. the systematically varied sets of stimuli. Post-processing is applied regularly to ensure comparable audio characteristics for play-back. This does not mean ‘core’ signal manipulation in order to smooth the signal or adjust single parameters to systematically vary the stimuli. It rather refers to e.g. concatenation of stimuli into pairs for A-B comparisons, or adjustment of overall sound pressure level to become better suited for perception test environments.

The final stimuli are played back to a larger number of naive listeners with a specific task to complete. This task usually comprises a rating or judging of the audio signal to obtain an assessment with respect to the research question, e.g. whether stimuli are perceptually acceptable in general, or whether certain single articulatory parameters exert an auditory influence on listeners. The participants are asked to explicate their intuitive reactions, and we emphasize that there are no ‘correct’ and ‘incorrect’ answers but that we are interested in their “gut feeling.” Finally, the collected perceptual data are statistically analyzed.

Throughout our experiments we do not explicitly ask about the intelligibility of an utterance. However, during the imitation step we pay attention that basic intelligibility is provided in a stimulus.[4] Since we focus on the paralinguistic dimensions of speech we simply want to ensure that the linguistic message is understood. Thus, if whole words are synthesized for perception tests, we provide their written form in addition to the audio signal to facilitate the understanding of the wording.

This is done due to known interactions between intelligibility and other properties of an utterance that we are interested in, e.g. naturalness or overall acceptability. For instance, Klatt (1987) remarks in his TTS review that intelligibility and naturalness are not the same but “appear to be fairly highly correlated” (p. 778) when comparing different TTS systems. Naturalness he describes as “a multi-dimensional subjective attribute that is not easy to quantify. Any of a large number of possible deficiencies can cause synthetic speech to sound unnatural to varying degrees.” (p. 778) Thus, posing very general questions such as “How natural does the utterance sound?” represents, in fact, a very complex task because naturalness interacts with, or is a result of, many facets of the voice. We nevertheless use this approach because in the end, it is the overall impression that counts for a listener in the real world. Additionally, with naive listeners, this is in our view the most practical way since the overall impression is what is easiest to retrieve without any particular phonetic training.

Finally, it should be noted that speech perception is not the same in every individual and may change intra-individually, due to situational variables such as noise or attention (see e.g. Fellbaum, 2012: 134 ff), or due to personal experience, learning, recency effects etc. (cf. e.g. explanations from Exemplar Theory, Pierrehumbert, 2001; Walsh et al., 2010). This may for instance affect perception of aged voices or regionally accented voices. We therefore collect demographic and situational data to be able to take into account possible individual influences during the analysis of the perceptual data.

Perceptual evaluation using the world-wide web

Most listening tests of this thesis are performed via the world-wide web by means of web-based test interfaces. Every participant sits in front of their personal computer (or other internet device) and takes the test on their own at a time they want. In our tests, we always design the user interface in such a way that one single stimulus is presented on one slide. After rating the stimulus, a participant proceeds to the next slide. The tests are available online for an extended period of time during which we invite people via email and newsletter postings to participate in the test and distribute the invitation further. In Appendix B, we

[4] Systematic intelligibility evaluation could e.g. be done with semantically unpredictable sentences (SUS, cf. e.g. Benoît et al., 1996, Picart et al., 2012), even using largely automated scoring procedures (Bunnell and Lilley, 2007).

provide background information on web-based listening tests and point to quality-ensuring measures when designing and implementing a test as well as when analyzing and reporting the data. In the following, we briefly describe the software that we used for our tests. Screenshots of the web-based perception tests in this thesis, including English translations of relevant textual elements can be found in Appendix Section B.5.

For the first tests, we used the software package WebExp 2.0 from CSTR (Keller et al., 2009). It runs as a Java applet in the participant's internet browser and is hosted on a web server on which the researcher administers the installation and the launch of the service after having adapted certain configuration files written in XML format. The tool kit was not entirely suited for all our purposes (see below) but it was a very useful working solution since implementing a test from scratch or using only locally accessible test software was not an option. WebExp 2.0 offers a configuration where the rating scales are activated only after the stimulus was clicked upon at least once for play-back. This prevents participants from 'blindly' clicking through the test without listening to the stimuli and is therefore a simple measure to increase the quality of the results. Another measure for quality improvement is to use reaction times as outlier criterion. Unfortunately, contrary to our expectations, the recorded reaction times in WebExp 2.0 turned out to be unsuitable to serve as outlier criterion during data analysis because of technical problems.[5]

For the regional accent test presented in Chapter 11, we used the software service PERCY from the Bavarian Archive for Speech Signals BAS (cf. e.g. Draxler, 2011).[6] As with WebExp 2.0, it can be configured in such a way that complete stimulus playback is obligatory prior to rating input. In addition, we were able to implement a specifically balanced test design in an easier way than would have been possible in WebExp. Lastly, we had access to reliable reaction times of the participants. So we were able to use the data as a post-experimental control whether a participant was focused on the task and not doing anything else in between. With PERCY, we recorded several time stamps including start of slide presentation, start of audio playback, stop of last audio playback, input of final rating, which automatically also advanced the test to the next slide (next stimulus).

4.2.8 Final articulatory interpretation

After the imitation and three-fold evaluation of the targeted speech phenomenon (articulatory, acoustic, perceptual), we integrate the partial results into a final summary and discussion, which includes an interpretation of the proposed articulation with regard to the research question. In other words, if the articulation seemed plausible from a technical perspective, and the acoustic and perceptual evaluations were successful, we translate the model articulatory suggestions back to human terms to obtain a description of possible articulatory details as suggested by the synthesizer. Please note again that the suggested articulation can only be

[5] The stimulus audio playback sometimes did not work straight away, so that the recorded times were not mirroring the real reaction times.

[6] http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

regarded as *one* possible way to produce a target phenomenon and that there may be others, just as humans may use different strategies to produce a given sound.

Each discussion may also touch on issues of how well grounded our particular evaluation can be, based on the notes above at the single steps. This includes e.g. a discussion of possible limitations and optimization possibilities regarding a) the stimuli and their articulatory plausibility, b) the evaluation and the reliability of the perceptual design, and c) the synthesizer.

4.3 Overview of the experiments

In this section, we present an overview of the topics of the experiments that we conducted with the articulatory synthesizer, using the method of imitation and evaluation outlined in Section 4.2 and Figure 4.1. After some general remarks, we present the experiments' articulatory-phonetic topics (Section 4.3.1) and briefly sketch the synthesis technical aspects of the experiments (Section 4.3.2). The experiments themselves are presented in detail in Chapters 5 to 11.

The topics of the experiments span a variety of phonetic phenomena, all pertaining to paralinguistic levels of speech communication. The link between them is that they, taken together, cover all major parts of the speech tract. This means that each experiment is used to investigate fine articulatory detail in different parts of the vocal apparatus, to a) find articulatory suggestions for given speech phenomena (phonetic research questions) and b) to evaluate the performance and suitability of the articulatory synthesizer in this area of speech production (technical research question).

We conducted the studies as a series of experiments moving from short, simple speech imitations to more complicated ones, i.e. the internal complexity and the length of the stimuli (in numbers of segments) increases through the course of the following chapters. We started out with short stimuli since they were easier to master and we could concentrate on very specific articulatory aspects. As the speech imitations become longer and more elaborate, other aspects such as coarticulation and prosody become increasingly important and require attention even if they are not always the center of the investigation.

4.3.1 Articulatory-phonetic aspects

An overview of the seven experiments highlighting their main phonetic (articulatory) research questions can be found in Figure 4.3, a concluding overview of the main phonetic findings is presented in Figure 12.1 in the final chapter (p. 194).

We begin with an experiment on larynx-height-associated voice quality in stationary vowels (monophthongs) because it represents the foundation of all the other experiments. In this study we test fine-grained acoustic requirements that are basic to all other experiments and test whether the synthesis software is at all detailed enough to capture some of the more subtle variations in speech production, such as variations in voice quality and

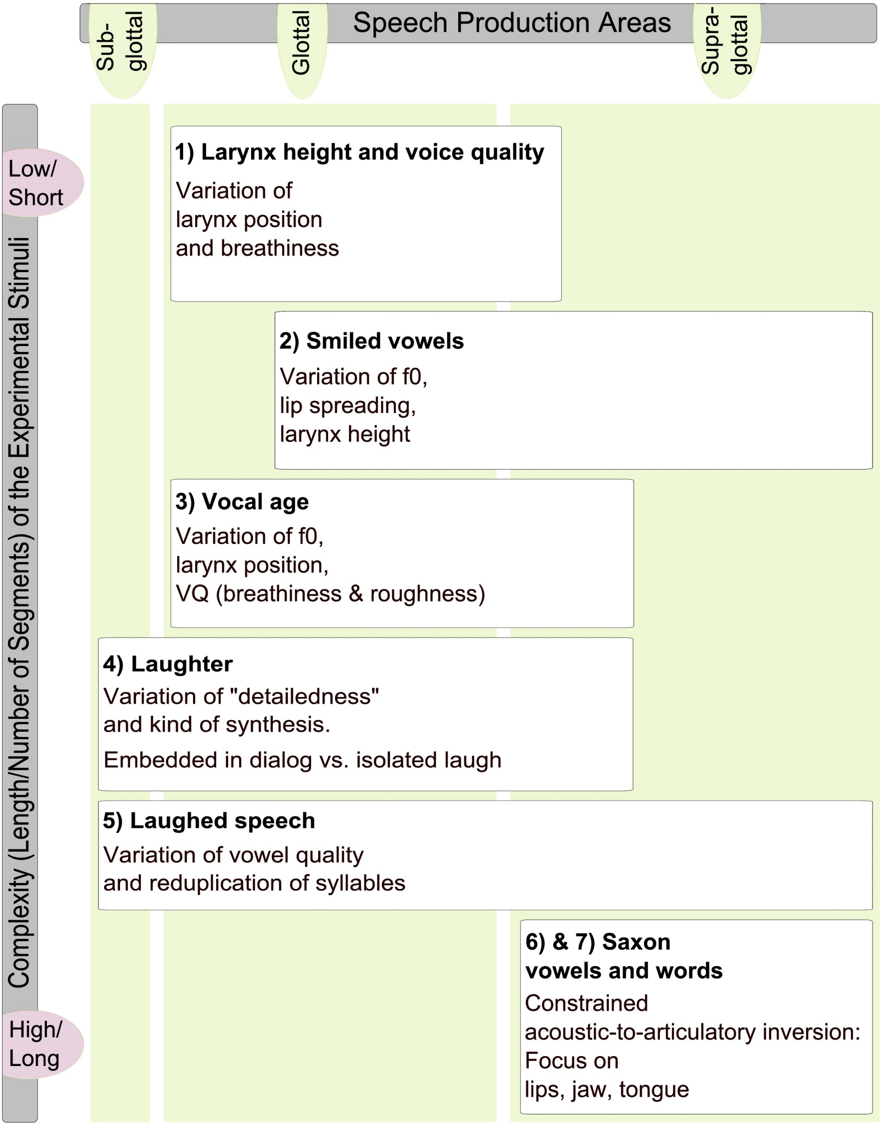

Speech Production Areas
Sub-glottal
Glottal
Supra-glottal
Complexity (Length/Number of Segments) of the Experimental Stimuli
Low/ Short
High/ Long
1) Larynx height and voice quality
Variation of larynx position and breathiness
2) Smiled vowels
Variation of f0, lip spreading, larynx height
3) Vocal age
Variation of f0, larynx position, VQ (breathiness & roughness)
4) Laughter
Variation of "detailedness" and kind of synthesis.
Embedded in dialog vs. isolated laugh
5) Laughed speech
Variation of vowel quality and reduplication of syllables
6) & 7) Saxon vowels and words
Constrained acoustic-to-articulatory inversion: Focus on lips, jaw, tongue

Figure 4.3: Overview of the seven experiments of this thesis: topic and test parameters; vertically arranged by increasing complexity of the stimuli (i.e. number of segments). The horizontal length of a box indicates to what extent the respective areas of speech production are covered. *VQ* = voice quality.

fine formant structure. Once this foundation is established, we move on to an experiment concerning smiled speech. Again, monophthongs are used as carriers for the stimuli, and this time laryngeal as well as supraglottal parameters are manipulated (lips position, larynx height), resulting in more or less 'smiley'-sounding utterances. A similar setup is used in the third experiment, where we aim at simulating young vs. old voices, each uttering different diphthongs. The different vocal ages are characterized mainly by, again, varying laryngeal settings resulting in different voice qualities, plus a slight supraglottal manipulation in the pharyngeal region of the vocal tract. The non-stationary nature of the diphthongs (vs. monophthongs) requires basic coarticulation mechanism to be applied here.

Even more coarticulation is used in the fourth experiment where a short word is imitated in a speech-laughed manner, i.e. laughing and speaking take place at the same time, entailing competing demands on the vocal apparatus. This experiment focuses on syllable reduplication which requires laryngeal processes that are not so common in 'regular' speech. Again, voice quality variation is also part of the study. In the fifth experiment, we exploit – and thus evaluate – the laryngeal manipulation possibilities to an even larger extent by imitating a stretch of laughter, which requires unusual actions at the vocal folds. Of additional special interest are the capabilities of the pulmonic component of the synthesizer, since laughter also puts unusual demands on the pulmonic air flow, including high pressure levels and very fast variation of these levels. Lastly, also supraglottal manipulations are used to generate the laugh vocalization, so this experiment is one of the more complex ones regarding the structure of the stimuli. It is also a quite challenging one since the articulatory demands for laughing are quite different from the ones for speaking.

With Experiments VI and VII we turn to 'regular' speech, with increasing stimulus complexity. The experiments aim to imitate two-syllable words spoken with a Saxon accent, and focus on the effects of supraglottal vocal tract manipulations. We use the restricted-scenario acoustic-to-articulatory inversion algorithm introduced in Section 4.2.4 on page 56 in order to create the desired regionally accented vowel imitations (Experiment VI). Additionally, in contrast to the vowel-stimuli in the first experiments, we integrate the vowels into short words (Experiment VII). Thus, a number of additional requirements have to be met in this study, such as proper coarticulation, timing and duration of acoustic segments, and word intonation, to obtain intelligible utterances.

Taken together, the experiments cover all areas of the vocal apparatus: Glottal, supraglottal, and subglottal. All studies except for the one on regional accents involve glottal parameters as a central part of the experiment, be it in terms of voice quality adjustments or distinct f_0 manipulations and extreme vocal fold movements as used for laughter imitation. Manipulations of the supraglottal settings (vocal tract shape) are central to all studies but the one on laughs. Changes in the vocal tract shape take place either at its extreme ends (larynx position and lip shape), or they involve all articulators together to determine the overall shape of the oral cavity. The study on laughter is the only experiment that features subglottal activities as a central variable, namely large changes of lung pressure.

Table 4.1: Relevant synthesis parameters in the experiments. For more details on the parameters please refer to Section 3.1. For an overview of the involvement of each parameter in the experiments, please refer to Figure 12.2 in Section 12.1.1. *GS* = gestural score.

Parameter name	**Where specified**	**Determines ...**	**Influences ...**
Parameters of the different speech production areas			
Vocal tract configuration (VT)	In config file; called by designator on GS (vocalic and consonantal tiers)	Vocal tract geometry for phonemes	Acoustic sound quality
Velum (vel)	Velic tier on GS	Size of velar opening	Degree of nasal coupling/nasality
Glottal (glot)	Glottal tier/f_0 tier on GS	Position of vocal folds; fundamental frequency	Voice quality; intonation
Pulmonic (pulm)	Pulmonic tier on GS	Level of air pressure from lungs	General volume of utterance; aspiration or burst intensity
Temporal and procedural (coarticulation) parameters			
Duration	As an attribute of each gesture on GS	Length of gesture	Length of acoustic segments, hypo-articulation
Vocal effort (slopes) and target values	As an attribute of each gesture on GS	Velocity of articulators towards target; degree to which VT target configuration is reached	Precision of articulation, hyper- and hypo-articulation
Gestural alignment	By placement of gesture on time axis of GS	Temporal coordination among any two gestures	Intelligibility of sounds, proper acoustic course of events in complex sounds (e.g. plosives)
Dominance values	In config file; called internally by dominance module	Relative importance (dominance) of each geometric parameter within a phoneme compared to another phoneme	Final articulatory trajectory if two or more gesture commands are competing

An additional study conducted with VocalTractLab is briefly discussed in Section 12.2. It deals with the perceived levels of uncertainty in answers in a fictitious human-machine-dialog. Similar to the word synthesis in Experiment VII (Chapter 11), we use VTL to synthesize short words and phrases as stimulus material. The focus however is not put primarily on articulatory aspects but on general linguistic and pragmatic research questions. Therefore this study is not part of the articulation-centered research goal of the thesis. However, it demonstrates how VTL can be used to investigate more general phonetic-linguistic phenomena of spoken communication, by conducting general linguistic experiments where the articulatory details are simply assumed to work in the background.

4.3.2 Synthesis technical aspects

From a synthesis technical perspective, the experiments each cover a range of parameters available in the articulatory synthesis software VTL. The parameters have been introduced in Section 3.1. Taken together, their behavior can be used to characterize the technical performance of the synthesizer regarding its capability to simulate articulatory processes. For an overview of these technical parameters and their influence on the speech sounds, please refer to the listing in Table 4.1. For a concluding overview of which experiment focused on which parameters, please refer to Figure 12.2 in the final chapter (p. 195).

The parameters can be grouped into two sets. The *speech production area parameters* comprise settings related to: the vocal tract shape, the position of the velum, the settings at the glottis and pulmonic settings, thus covering supraglottal, glottal, and subglottal areas of articulation. The *temporal and procedural parameters* comprise aspects of segmental duration, gestural slopes and target values, gestural alignment and dominance values. The first group of the experiments (Chapters 5, 6, 7) deal with short vowel stimuli and therefore cover mainly the ‘static’ parameters of the different speech production areas and not so much the temporal and procedural parameters. The second group of the experiments (Chapters 8, 9, 10/11) also cover the dynamic parameters because their stimuli are more complex and demand more elaborate gestural scores.

After having presented in this chapter the main methodological aspects common to all experiments, including assumptions and goals as well as procedural details, in the following chapters we present the individual experiments sketched above, including introductory sections with theoretical background information on each of the topics covered. After presentation of the last experiment, we summarize the main findings and bring together the individual strands that have been pursued in the experiments. The thesis concludes with a general discussion and outlook.

Chapter 5

Experiment I – Larynx height and voice quality

In this chapter, we present a comparative study of natural and synthetic speech samples that vary in larynx height. The acoustics of isolated vowels with different vertical positions of the larynx are analyzed with regard to subtle changes in the properties of the voice. In terms of the overview of all the experiments presented in Section 4.3, we deal with short stimuli, i.e. individual vowels and focus locally on the glottal area, and also on the lower end of the supraglottal area since larynx height variation influences the length of the pharynx.[1]

The speech production goal of this study is to test the hypothesis that different vertical larynx positions should be accompanied by certain voice qualities in order to reflect natural voice quality properties. Laver (1980) describes voice quality as "the characteristic auditory colouring of an individual speaker's voice" (p. 1). He presents a model-theoretic 'neutral setting' of the speech organs and compares other settings to it. Our focus here is on the laryngeal setting of breathiness, and we apply different degrees of breathiness to different vertical positions of the larynx.

The technical evaluation goal of this study is twofold. One aspect is to assess the capabilities of the synthesis software in terms of voice quality, the other aspect concerns the fine formant frequency structure. Voice quality in general is a core factor in developing natural and adequate sounding synthetic speech. Thus simulating the desired properties of voice quality associated with larynx height would be a great asset of the synthesizer in terms of paralinguistic quality.

The fine formant structure of vowels deals with the small differences in formant values while keeping the overall quality of a vowel unchanged. For a listener to perceive an /aː/ for example, the resonance frequencies of the vocal tract can be within a certain range of values. Different speakers use slightly different typical formant frequencies for a given vowel quality

[1] These empirical data have been published in Lasarcyk (2007).

and this is partly due to the length of their vocal tract. The variability in the acoustic domain is also due to slightly different articulation strategies but we will focus on the vocal tract length here. The length usually varies between speakers due to anatomical differences but it can also vary within a speaker when they voluntarily change vocal tract length as a means to express themselves. This may e.g. be the case when they intend to sound larger or smaller by temporarily lengthening or shortening the vocal tract (frequency code, Ohala, 1983). In some languages, larynx height is also used as part of register (e.g. Asian languages like Yi and Bai, Edmondson et al., 2001). Apart from that, changes of vertical larynx position also seem to be involved in smiled speech (see Chapter 6). With this study, we aim to assess how well acoustic properties are imitated regarding the fine formant structure of vowels when the vocal tract length changes.

We present the major mechanisms of speech production that are used to move the larynx up and down, and the effects of this movement on the spectral structure and excitation quality of the voice (Section 5.1). We then describe the procedure of human speech data collection and analysis as well as the generation of the synthetic speech data (Section 5.2), present the results of the comparative speech data analysis for voice quality parameters and fine formant structure (Section 5.3), and discuss the implications for human speech production as well as the implications of this study for the usability of the synthesis software (Section 5.4).

5.1 Larynx height in speech production

5.1.1 Human articulation

In this experiment, we study the acoustic correlates of larynx height in human speech production in three larynx height settings: raised, neutral and lowered. We then compare them to articulatory synthesized versions, paying special attention to synthesis parameters such as breathiness. Larynx height in a strict sense means vertical position of the larynx. The human speech production system, however, is a very complex and interdependent system of muscles, ligaments and tissues. Thus, changing the position of the larynx entails changes of other parts in the vocal tract as well, as we will see below. These include the state of the vocal folds and the shape of the vocal tract.

The main actor for modifying larynx height is the hyoid bone (Laver, 1980: 24ff). This bone is an interaction point of several muscular systems relevant to speech production. It can be assumed that a change in larynx height will cause changes in the surrounding body area. Moreover, while the infrahyoid muscles are pulling down the larynx, the suprahyoid muscles should relax to enable that movement. Observations suggest that lax voice, enabled by relaxed muscles, is often accompanied by a slightly breathy voice quality. Conversely, this relaxation is also found for the production of breathy voice (Laver, 1980: 31). So our assumption here is that the lower the larynx position, the more relaxed is the voice.

Another major effect of larynx lowering is the lengthening of the vocal tract. Vocal tract length is measured from the lips to the glottis. Thus, when the larynx is lowered, the glottis is moved further away from the lips, thereby increasing the length of the vocal tract.

5.1.2 Acoustics of larynx height

The acoustic changes in the voice that occur when the position of the larynx is varied can be relatively large. To some extent, one single speaker might sound like multiple speakers, depending on which larynx position they choose to use. Eckert and Laver (1994), for instance, give demonstrations in their audio samples of how variable voices can be. To provide a quantitative description of the acoustic changes that occur when a speaker varies larynx height, we present the analysis of a small sample of natural speech in Section 5.2. The results reflect the general findings in the literature which are described here.

Firstly, we find changes in formant frequencies. As mentioned in Section 5.1.1, varying larynx height changes the overall length of the vocal tract. From the acoustic theory of speech production (Fant, 1960) it is known that when the vocal tract length is changed, the formant frequencies of a given vowel will also change slightly while the overall vowel quality can be kept constant (Laver, 1980). Compared to speech in a neutral setting, we should find:

- Higher formant frequencies in a raised-larynx setting
- Lower formant frequencies in a lowered-larynx setting

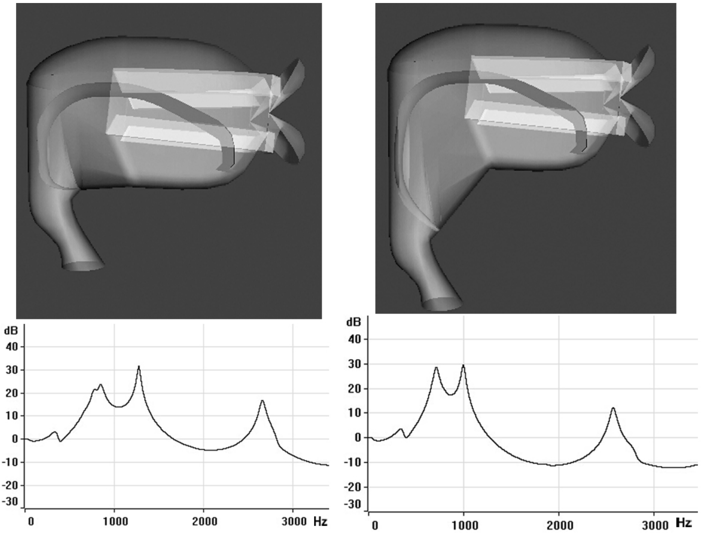
Apart from the influence on formant frequencies, which can be calculated directly from the vocal tract shape, natural speech is also influenced in terms of voice quality, as has been described in Section 5.1.1. The substantial vertical shift of the larynx affects the mode of vibration of the vocal folds (Laver, 1980; Strik and Boves, 1992; Sundberg and Askenfelt, 1981), leading to the following assumptions, compared to speech in a neutral setting (modal voice) when the larynx is between the two extreme positions:

- Raised-larynx speech sounds tenser
- Lowered-larynx speech sounds laxer

In Section 5.2.3, we present specific measurements used to distinguish different voice qualities.

5.1.3 Transfer into synthesis features

From the articulatory descriptions for larynx height and its acoustic implications for voice quality and formant structure, we derive two manipulation components to be implemented in the synthesis software to produce larynx-height adjusted speech samples.



(a) Raised-larynx setting.

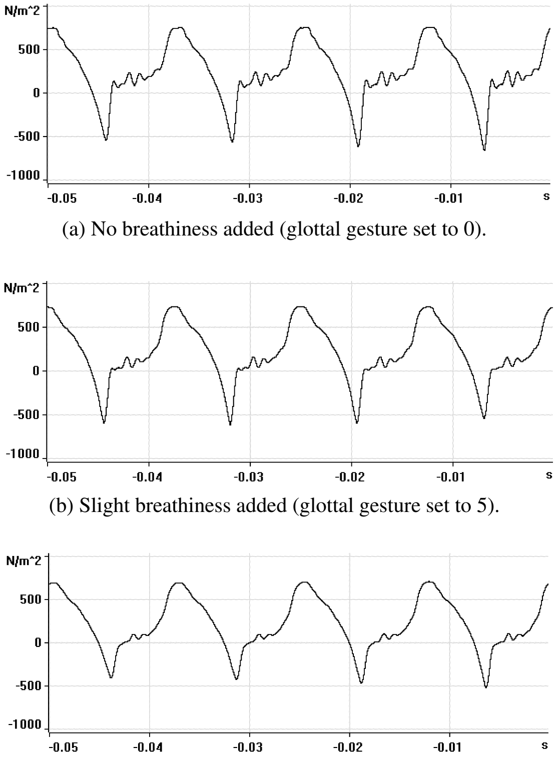
(b) Lowered-larynx setting.

Figure 5.1: Vocal tract configurations (top) and transfer functions (bottom) for /aː/ in two larynx height settings. Note the shape differences in the pharynx and the formant frequency differences especially in the first and second formant.

The first component is a variable describing larynx height. The larynx itself is not part of the three-dimensional model but the effects it has on the vocal tract shape can be controlled by using vocal tract shape parameters (cf. Section 3.1). In the case of larynx height control, we manipulate the parameter for the vertical tongue root position (HY). It accounts for changes of larynx height because the vocal tract is lengthened at the lower end when this parameter is lowered (see illustration for /aː/ in Figure 5.1).

As indicated in Section 5.1.2, we define three levels of larynx height settings: raised, lowered, neutral. This technical definition only includes the parameters of the vocal tract shape and no voice quality aspects. The neutral larynx height setting is taken from the VTL default configuration for each vowel, the raised setting is derived from these configurations by changing the vertical tongue root (HY) parameter to its maximum value (-3.50), and the lowered setting is derived by changing the parameter to its minimum value (-6.00, cf. Table 3.1, p. 29).

Changes in formant frequencies should not have to be added explicitly to the synthesizer since the moving down of the larynx variable HY automatically entails a lengthening of the vocal tract. It is of interest, however, whether these vocal tract manipulations do indeed automatically show human-like formant changes or not. Thus, the speech samples are analyzed with respect to changes in formant frequencies.



(a) No breathiness added (glottal gesture set to 0).

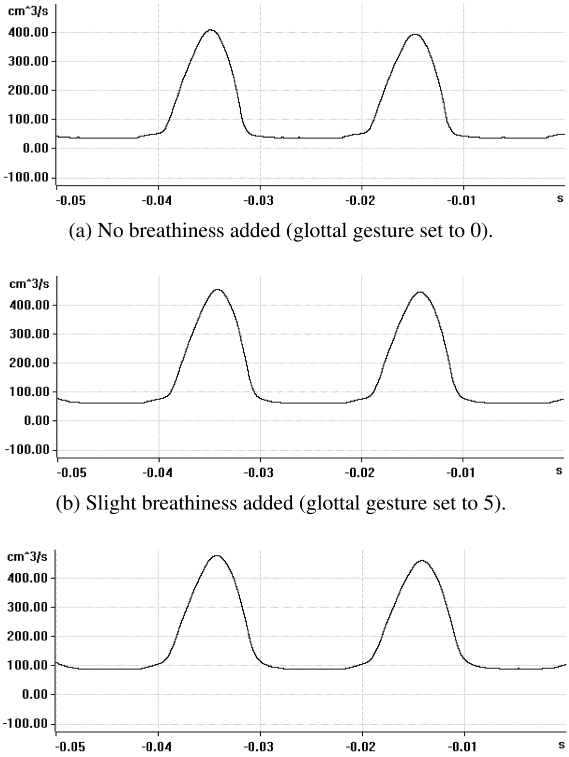
(b) Slight breathiness added (glottal gesture set to 5).

(c) Moderate breathiness added (glottal gesture set to 10).

Figure 5.2: Pressure at glottis during phonation of [aː] with different degrees of breathiness, $f_0 = 83\,\text{Hz}$. For details see Section 5.1.3.

The vocal tract images in Figure 5.1 illustrate the changes of the vocal tract shape for the vowel /aː/. The main difference to be noted is located in the pharynx. For the lowered-larynx setting (b), the pharyngeal part of the oral cavity is extended, leading to an overall lengthening of the vocal tract. Figure 5.1 (lower half) shows the corresponding vocal tract transfer functions. The peaks in the spectrum denote the positions of the formants for that vowel. For the raised-larynx setting (a), the formants are raised (F_1 around 850 Hz, F_2 around 1250 Hz), whereas the lowered-larynx setting (b) shows lowered formants (F_1 around 750 Hz, F_2 around 1000 Hz).

The second manipulation component is voice quality, mainly in terms of breathiness as motivated in the previous section. The voice-quality effects of the larynx-height changes in humans let us assume that it is necessary to change more than the larynx position variable in order to obtain ‘human-like’ results with the synthesizer. Thus, we vary the degree of



(a) No breathiness added (glottal gesture set to 0).

(b) Slight breathiness added (glottal gesture set to 5).

(c) Moderate breathiness added (glottal gesture set to 10).

Figure 5.3: Volume velocity at glottis during phonation of [aː] with different degrees of breathiness. For details see Section 5.1.3.

breathiness by adjusting the default position of the vocal folds to different degrees of abduction. Different amplitude values for the glottal abduction gesture (“open”) are used, where a higher number means greater degree of abduction (cf. Section 3.1 for technical details on the synthesis system).

For this experiment, we define three degrees of breathiness to add to the different levels of vertical larynx position: none, slight and moderate. Auditorily, they correspond to tense, modal and lax voice, respectively. The system’s scale extends from 0 to 100, with 100 equaling complete abduction, i.e. vocal folds being pulled apart completely. Based on auditory assessment, we define 0 as the vocal fold position for tense voice (i.e. no breathiness), 5 for modal voice (slight breathiness) and 10 for lax voice (moderate breathiness). Later on, all three levels of breathiness will be combined with all three levels of vertical larynx position.

Figures 5.2 and 5.3 depict the details of the excitation signal in terms of pressure and volume velocity at the glottis. With increasing breathiness, the pressure difference from minimum to maximum decreases and the excitation function becomes smoother. This results in a greater attenuation of the higher frequency components; the spectrum falls off more steeply, i.e. with a greater energy decrease per octave. At the same time, volume velocity at the glottis increases when breathiness increases. This applies to the open as well as the closed phase of the glottal cycle, indicating that a glottal leak is letting more air pass through, as can be presumed for vocal folds that are being kept apart at a certain distance to generate breathy voice.

5.2 Data and analysis

5.2.1 Human speech data

The human speech data are recorded from an adult male speaker who is able, by training, to control his manner of articulation. His task is to produce isolated vowels while focusing on larynx height control. Checks for correct larynx height are performed visually and auditorily. Other vocal tract properties that are needed to produce a given vowel are kept as constant as possible. Despite these intentions for invariant articulation, it is assumed that some features in the speech production process still change involuntarily. It is these changes that we want to capture in our analysis. We record the three vowels /aː iː uː/, each produced multiple times in a raised, neutral and lowered larynx setting. We select for analysis those vowel tokens where speaker and experimenter agree that the larynx setting has been optimal. The number of tokens can be found in Table 5.1.

Table 5.1: Number of human speech vowel tokens at each larynx height selected for analysis.

Setting	/aː/	/iː/	/uː/
Raised	2	5	5
Neutral	2	3	3
Lowered	1	2	3

5.2.2 Synthetic speech data

For each vowel /aː iː uː/, we synthesize 9 different simulations by combining each larynx height with each degree of breathiness (3x3=9) as depicted in Table 5.2. The /aː/, for instance, is synthesized in the raised-larynx setting with no, slight, and moderate degree of breathiness. The other two larynx settings are also combined with each of the three voice qualities.

In this set of 9 simulations for each vowel, we assume that only one triple of them resembles the human speech production properties in terms of voice quality. It is the triple

Table 5.2: Matrix of voice 'profiles' of the vowel simulations generated synthetically by combining different degrees of breathiness (none (0), slight (5), moderate (10)) with each of the three vertical larynx positions (raised (R), neutral (N), lowered (L)). The cells written in bold face indicate the triple with the hypothesized 'human-like' combination of these two manipulation parameters (cf. Section 5.2.2).

Vertical larynx position	Degree of breathiness		
	None	Slight	Moderate
Raised	**R0**	R5	R10
Neutral	N0	**N5**	N10
Lowered	L0	L5	**L10**

in which breathiness is increased as larynx height is decreased, reflecting a laxer voice when the suprahyoid muscles relax to allow for the larynx to be pulled down (cf. Section 5.1.1). The corresponding designators are written in bold face in Table 5.2, and we will call this configuration the 'human-like' combination of manipulation components.

5.2.3 Analysis

Both the human and synthetic vowel tokens were subjected to the same analysis procedure and the results are interpreted in a relative fashion: The measurements are always compared over a complete set of three larynx settings (a triple) and the relative change is noted.

Spectral analysis

Using Praat (Boersma, 2001), we analyze the following in a stationary part of each vowel:

- Formant frequencies of the first three formants (F_1, F_2, F_3) in Hz
- Spectral energy of the first two harmonics (H_1, H_2) and the first three formants (A_1, A_2, A_3) in dB

The formant frequency values are extracted automatically and averaged over the selected stationary part of the vowel. To obtain the values of spectral energy, we visually inspect the amplitude spectrum (FFT) of a spectral slice spanning that same stationary part of the vowel. No smoothing is applied to the spectral slice. The spectral energy of H_1 and H_2 is read directly from the amplitude spectrum, while the amplitudes of the formants are read from the value of the closest harmonic.

Calculation of voice quality measurements

While the formant frequency values can be used directly to describe larynx height influences (see next section), the following four voice quality measurements are derived from the amplitude values in additional steps of calculation. Claßen et al. (1998) present these voice quality measurements to distinguish stressed from unstressed speech. Since stressed speech means a higher articulation effort, we assume that stressed vowels are somewhat tenser and unstressed vowels rather lax. Under this assumption, we apply these measurements to distinguish between tense and lax voice quality here.

- $H_1 - H_2$: Amplitude differences of H_1 vs. H_2 in dB; the time in which the glottis is open per glottal cycle (open quotient).
- Amplitude differences between H_1 and the first three formants, normalized in dB per octave:
  - $H_1 - A_1$: Degree of glottal opening during the whole oscillation period of the vocal folds.
  - $H_1 - A_2$: How abruptly the glottis is being closed (skewness). A gradual cutoff leads to a greater loss of energy in the higher frequencies.
  - $H_1 - A_3$: Velocity at which the airflow is cut off (rate of closure).

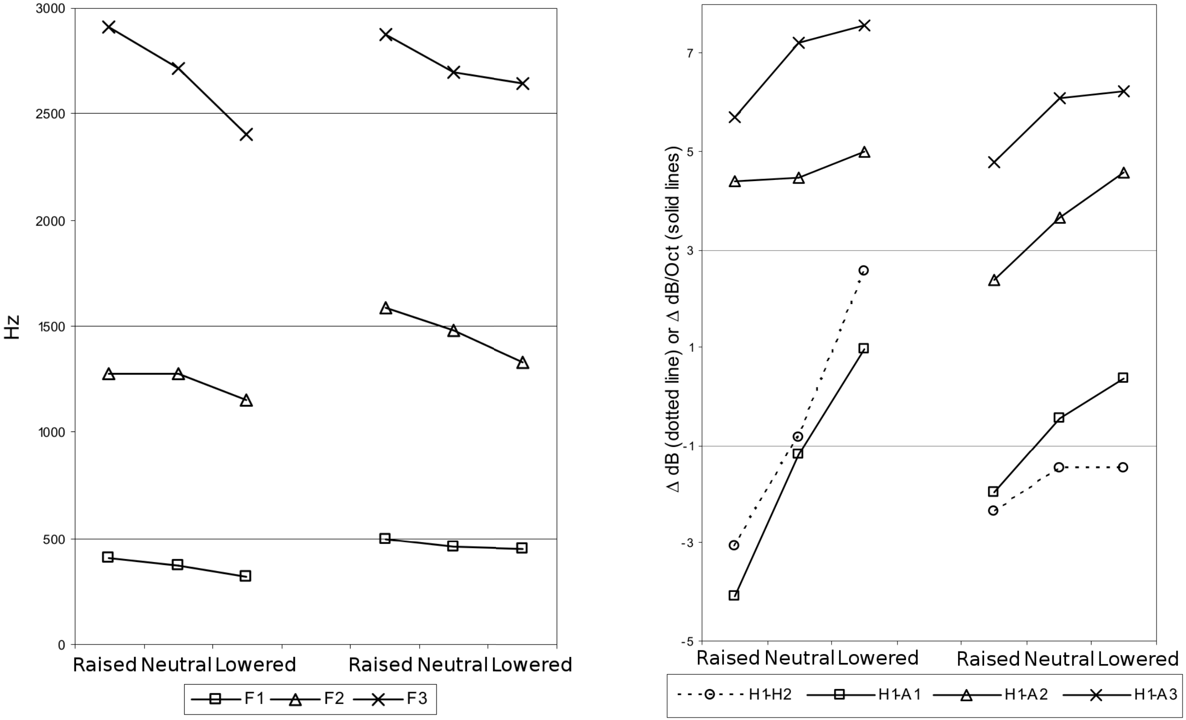
We calculate and average these voice quality measurements for each vowel, and then average over all vowels. In summary, higher amplitude differences indicate a laxer voice quality and can also be associated with a lowered larynx setting.

5.3 Results

Both the human and the synthetic speech samples show a lowering of formants during vocal tract lengthening, i.e. when the vertical position of the larynx decreases (see Figure 5.4a). This supports the hypothesis introduced above that a longer vocal tract in the synthesizer automatically causes lower formant frequencies for a given vowel, comparable to human behavior. It has to be noted, however, that the results for /uː/ (not depicted) diverge from this pattern when looking at the changes from neutral to lowered larynx. This is because the neutral configuration of /uː/ already uses the minimal HY value by default. When applying the manipulation to obtain the vocal tract shape of the lowered-larynx setting – setting HY to its minimum value – the shape does not change. Therefore, the formants do not change from neutral to lowered larynx settings.

Regarding voice quality (Figure 5.4b), the human data samples indeed show an indication that the elevation of the larynx is accompanied by an increase of phonatory tension (tense voice). The tension is reflected by a flatter spectrum for raised larynx voice, i.e. the higher frequencies are less attenuated than for lax voice. Thus, the voice quality measurements reveal smaller (difference) values for tense voice than for lax voice. In Figure 5.4b this is shown by rising values across the triple of larynx height settings from raised to lowered larynx.

The same tendency as for the human data can be observed in that particular subset of the synthetic speech samples which was assumed to most accurately mirror the human production process. Thus, when the lowest larynx setting is combined with the highest degree



(a) Formant frequencies of vowel stimuli of /aː/ over the three larynx height settings *Raised, Neutral, Lowered*. Left: natural speech; right: synthetic speech.

(b) Voice quality of vowel stimuli over the three larynx height settings *Raised, Neutral, Lowered*. Left: Natural speech; right: Synthetic speech. The synthetic samples depicted here contain the 'human-like' combination of larynx height with breathiness (cf. Table 5.2).

Figure 5.4

of breathiness, and then breathiness decreased as larynx height increased (cf. Table 5.2), the synthetic voice quality measurements point in the same direction as the human ones (cf. Figure 5.4b, right half). When this particular 'human-like' combination of larynx height with breathiness is not applied, the voice quality measurements are different and do not line up as nicely over the three larynx height settings as in the triple in Figure 5.4b. For reference, these results are shown in Appendix Figures C.1 to C.2.

5.4 Discussion

We first discuss the phonetic findings (Section 5.4.1) before concluding with a first assessment of the features of VTL (Section 5.4.2).

5.4.1 Phonetic aspects

In analyzing a set of vowels produced with different vertical larynx positions by a human speaker, we could confirm the general effects that larynx height has on voice quality. A lowered larynx position brought about voice quality measurements describing a laxer voice whereas an increase of larynx position led to an increase of phonatory tension. Having proposed that this tension could be closely related to the degree of breathiness of the voice, we designed synthetic vowels accordingly and found that, of any triple combination of larynx height with breathiness, only the one that follows this proposed strategy showed meaningful (i.e. 'human-like') voice quality measurements.

In terms of the speech production goal of this experiment, the results can be seen as an indication that, indeed, in human speech production, lowered-larynx voice is laxer and accompanied by more breathiness than neutral and raised larynx conditions. Other combinations seem a lot less natural and therefore arguably improbable.

5.4.2 Synthesis technical aspects

In terms of the technical evaluation goal of this experiment, the results indicate that the synthesis software has the capacity to simulate subtle influences of vocal tract shape and excitation quality: The fine formant structure related to vocal tract length reflected the changes that occur during human speech production and we were able to imitate 'human-like' voice quality properties.

When using the synthesizer in basic speech research this means that we can exploit the independence of sub-systems (such as vocal tract shape/larynx height and glottal state) to check hypotheses about speech production. Knowing that the synthesis software is capable of reproducing near-human voice characteristics in the above experiment, we can adopt this analysis-by-synthesis paradigm to other research questions. A hypothesis is confirmed when the synthesis software gives best (most 'human-like') results for the hypothesized 'human-like' way of producing speech and gives clearly worse results when opposite (less 'human-like') strategies are used.

In all cases, however, the analysis-by-synthesis method should only be regarded as a complementary approach to refine our knowledge about speech production. The initial hypotheses will always be derived from observations of real speech production, be it by introspection, visual inspection, or by means of articulatory data acquisition as described in Chapter 3.

Chapter 6

Experiment II – Smiled vowels

In this chapter, we present an articulatory experiment involving the production and perception of smiled speech. A listener can distinguish smiled from non-smiled speech even without being able to see the speaker's face. Based on the idea that the acoustic impression of smiled speech is correlated with a shortened vocal tract, we create synthetic smiled vowels that feature retracted lips and raised larynx. In a perception test, we explore the relative contributions of these two vocal-tract shortening features, combined with a rise in fundamental frequency.[1]

Thus, the speech production goal of this experiment is to test the hypothesis that spread lips, raised larynx and increased fundamental frequency all contribute to the 'smileyness' perception in the vowels /iː aː uː yː/.

The technical evaluation goal of this experiment is to further assess how authentically the synthesizer can replicate the fine formant structure of a shortened vocal tract when not only the larynx is raised (as in Chapter 5) but also the lips are spread.

With regard to the overview of all experiments (Section 4.3), this chapter mainly involves supraglottal manipulations regarding vocal tract shape and length. It also deals with glottal aspects of speech production because the test stimuli involve f_0 manipulation.

After presenting background information about the characteristics of natural smiled speech, we derive the articulatory parameters for this experiment (Section 6.1). Based on these parameters, synthetic smiled speech data are generated and evaluated (Sections 6.2 and 6.3). The chapter concludes with a discussion, raising articulatory as well as technical considerations related to the experiment (Section 6.4).

[1] These empirical data have been published in Lasarcyk and Trouvain (2008).

6.1 Smiling during speech production

6.1.1 Articulation and acoustics

Several studies report that smiled speech can be distinguished auditorily from non-smiled speech (cf. e.g. Tartter, 1980; Tartter and Brown, 1994; Schröder et al., 1998; Robson and MackenzieBeck, 1999; Drahota et al., 2008). Parameters which were found to be typical of smiled speech comprise raised f_0 and raised formant frequency values (Tartter, 1980; Tartter and Brown, 1994; Robson and MackenzieBeck, 1999). Increased values for F_1 and F_2 can be explained with a shortening of the vocal tract that occurs when the corners of the mouth are retracted for smiling. This "i-face" in Ohala's frequency code (Ohala, 1983) – in contrast to the "o-face" – is also suggested as a typical setting for signaling smallness of the speaker (cf. Xu and Chuenwattanapranithi, 2007) by increased formant frequencies, indicating a smaller vocal tract, and increased phonation rate, indicating smaller vocal folds.

Less well explored (but cf. Xu and Chuenwattanapranithi, 2007) is the possibility of shortening the vocal tract by raising the larynx (cf. Figure 6.1), as has been observed for varying the vocal tract length during vowel production (Perkell, 1969). This can have an effect of a) raising the formant values (cf. Chapter 5, see also Xu and Chuenwattanapranithi, 2007), and b) raising f_0 as used in some Asian languages as part of their register system (cf. Edmondson et al., 2001).

This experiment seeks to find the relative contributions of the three factors lips, larynx, and f_0, which are possibly responsible for the perceptual effect of 'smileyness' in speech. In a similar but not identical study, Xu and Chuenwattanapranithi (2007) showed that a manipulation of these three parameters could be used as cues for body size and anger–joy distinction.

The parameters are manipulated individually, using the articulatory synthesizer VTL. With human speakers, in contrast to the articulatory synthesis approach, several methodological problems would be expected: Natural smiled speech occurs as a holistic impression, making it difficult to separate the individual articulatory factors. In addition, speakers constantly vary the intensity of smiling (Schröder et al., 1998; Drahota et al., 2008), which is also observable in the degree of lip spreading (Robson and MackenzieBeck, 1999). Furthermore, the effects of genuine ('felt') smiles and artificial ('non-felt') smiles on speech are still unclear (cf. Schröder et al., 1998; Drahota et al., 2008). Finally, measuring larynx height is not a straight-forward procedure (cf. e.g. Fagel et al., 2009).

6.1.2 Transfer into synthesis features

Motivated by the studies on smiling which involved human subjects, we focus on the following articulatory parameters in this experiment. They involve the position of the lips and the larynx, and the level of fundamental frequency.

The position of the lips is controlled by manipulating the degree of lip protrusion (*LP*, cf. Table 3.1, p. 29). This manipulation operates only in the horizontal plane and does not involve vertical distance changes between the lips, measured in the midsagittal plane. The position of the larynx is controlled by the vertical position of the hyoid bone (*HY*, see also Table 3.1), which is closely connected to the vertical positioning of the larynx, as described in Section 3.1.1 (p. 30). Based on the findings in the previous experiment (Chapter 5, p. 75), the change in larynx height is accompanied by corresponding changes in voice quality, involving varying degrees of breathiness. This is achieved by placing the corresponding glottal abduction gestures on the gestural score. The fundamental frequency is controlled on the f_0 tier by direct input of the target frequency values.

6.2 Data and analysis

Having selected the relevant articulatory parameters in this experiment, we describe the generation of the stimuli (Section 6.2.1) and present the methods by which they are evaluated: an acoustic analysis and transcription (Section 6.2.2), and a listening test (Section 6.2.3).

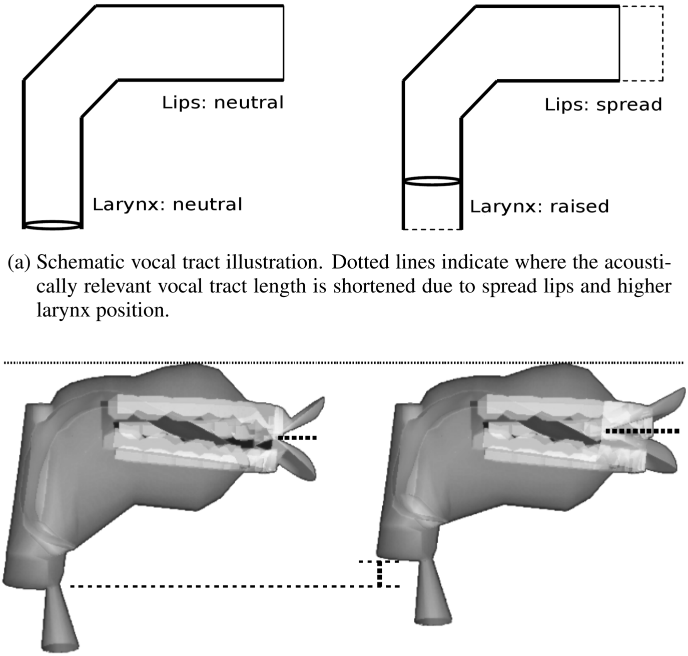
6.2.1 Synthetic smiled speech data

The four German vowels /iː aː uː yː/ are used for this experiment. The first three are chosen because they represent extreme points in the vowel space. The vowel /yː/ is selected as a candidate of front and rounded vowels. It will be of interest to compare its behavior to the *back* and rounded vowel /uː/ since lip spreading is involved in the experiment, which has similar acoustic effects as fronting.

A set of ‘neutral’ vowels is generated as a baseline condition. Each ‘neutral’ stimulus consists of a single vowel utterance with a duration of 560 ms. Its vocal tract target configuration is taken from the standard phone set of VTL (cf. Section 3.1.1, p. 28). f_0 is set at a monotonous 112 Hz across the entire vowel to avoid interactions with specific intonation contours which might possibly express some negative emotion. In a pilot test, the default intonation contour in simple vowel synthesis (rise-fall, in FDS mode, cf. Section 3.1.6, p. 39) caused associations with disgust. Of course, an effect of intonation can still not be excluded, and a monotonous f_0 contour might show a stronger tendency to be judged as ‘sad’ or ‘bored’.

To each ‘neutral’ or unmarked vowel, the following three changes are applied. This yields the corresponding ‘smiled’ vowel:

1. Spreading of the lips: The lips are retracted to the most extreme position possible.
2. Raising of the larynx: The larynx is placed in its highest position possible, combined with a slightly tenser voice quality, based on findings in Chapter 5.
3. Increasing of f_0: Fundamental frequency is increased by 2 semitones, i.e. from 111 to 125 Hz. The VTL default value of f_0 is taken as reference for the ‘neutral’ level of f_0.



(a) Schematic vocal tract illustration. Dotted lines indicate where the acoustically relevant vocal tract length is shortened due to spread lips and higher larynx position.

(b) Illustration for the base vowel /aː/ in VocalTractLab. Dotted lines are inserted for visual reference, to compare the different lip settings and larynx positions.

Figure 6.1: Illustrations of vocal tract shortening, induced both at the larynx and the lip end of the vocal tract. Left: neutral, right: shortened (smiled).

Illustrations of a ‘neutral’ vs. a ‘smiled’ vocal tract, with spread lips and raised larynx configuration, are presented in Figure 6.1. Besides the completely ‘neutral’ and the completely ‘smiled’ configurations, intermediate combinations are generated. In those stimuli, only one or two of the ‘smileyness’ parameters are active. The combination of all parameters results in 32 different stimulus vowels (4 vowels x 2 lip x 2 larynx x 2 f_0 settings = 32). An overview of the features of the stimuli is presented in Table 6.1.

6.2.2 Acoustic analysis and transcription of the synthetic smiled vowels

Since we use an articulatory speech synthesizer in a novel way for smiled speech, information on its performance is provided by measuring the acoustic vowel quality in the different ‘smileyness’ conditions. The formant analysis is done with Wavesurfer (Version 1.8.5, 2005; Sjölander and Beskow, 2000).

Additionally, the perceived vowel quality of each stimulus is transcribed. Due to the applied manipulations, one would expect possible confusions in the perception of vowel

Table 6.1: Cues of ‘smileyness’. NNN = neutral setting, SRH = complete ‘smileyness’ setting. Intermediate combinations are coded accordingly, with the first letter for coding the lip setting: neutral or spread (N/S), second larynx setting: neutral or raised (N/R), third f_0 setting: neutral or high (N/H).

Degree of ‘smileyness’	Lips	Larynx	Level of f_0
NNN	Neutral	Neutral	Neutral
NNH	Neutral	Neutral	High
NRN	Neutral	Raised	Neutral
NRH	Neutral	Raised	High
SNN	Spread	Neutral	Neutral
SNH	Spread	Neutral	High
SRN	Spread	Raised	Neutral
SRH	Spread	Raised	High

categories, since in contrast to Xu and Chuenwattanapranithi (2007), we do *not* manipulate further articulatory parameters to preserve vowel quality. Thus, six phonetically trained subjects marked which vowel quality they perceived for each stimulus on the IPA vowel chart.

6.2.3 Perception test design

In a formal perception test, subjects were asked to rate the perceived ‘smileyness’ of the 32 vowel stimuli. The perception test was carried out as a web-based experiment, using the framework WebExp 2.0 (Keller et al., 2009). The subjects were invited by email. Screenshots of the perception test slides can be found in Appendix Figures B.1 and B.2, p. 219.

The experiment started with an explicit warm-up phase: It did not serve as guidance on how to give answers, nor were any answers saved from this phase. It only familiarized the participants with the range of stimuli, layout, and technical process of the test setup (see also Section 4.2.7 and Appendix B for more information on web-based testing).

In the main experiment, the stimuli were presented in three blocks in randomized order. Using their home computer loudspeakers, 36 German speaking subjects participated in the experiment. They rated the stimuli on a five-point scale: “1” representing a vowel produced with “corners of the mouth pulled down”, “3” representing a ‘neutral’ setting, and “5” for “corners of the mouth pulled up.” As a visual shortcut, emoticons were used on the rating scale: “1” with the symbol ☹, “3” with 😐, and “5” with ☺. We cannot exclude associations with emotions, although we avoided giving direct hints to emotional states by mentioning terms such as sadness or happiness.

6.3 Results

In the following, we present the evaluation of the smiled vowel stimuli. Section 6.3.1 deals with the acoustic characteristics and transcription results, in Section 6.3.2 we present the results of the formal perception test.

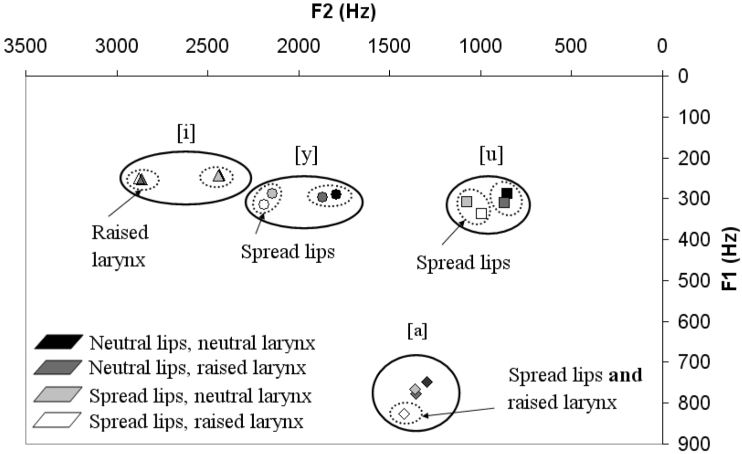


Figure 6.2: Formant plot of all stimuli in high f_0 setting. Both lip and larynx parameters influence the formant frequencies.

6.3.1 Acoustics and transcription of synthetic smiled vowels

Formant analysis reveals no relevant changes of formant frequency values when only f_0 is manipulated. As expected, however, lip spreading contributes noticeably to formant changes as does larynx raising (cf. Figure 6.2). For the two rounded vowels /uː yː/, spreading of the lips, which resembles a shortening of the vocal tract at the front end, raises mainly F_2. In the vowel /iː/, and to a lesser extent also in the vowel /aː/, only the raising of the larynx really contributes to a formant value increase. Here, the vocal tract cannot be shortened anymore at the lip end but the shortening takes place at the larynx.

The transcription task reveals that the perception of vowel quality is most stable across all variants of /iː/. For the different /aː/ stimuli, a slight fronting is perceived, resulting from spreading of the lips, raising of the larynx, or both. For /uː/, raising of the larynx results in a stable categorization of /uː/. However, lip spreading by itself, and both lip spreading and larynx raising lead to unstable perception results. Compared to that, the /yː/ variants induce an unstable perception even in the ‘neutral’ form, and lip *or* larynx manipulations lead to even more unstable category perceptions. Lip spreading *and* larynx raising stabilize the categorical perception to some extent. However, the perceived sound quality for this /yː/ is exclusively linked to *un*rounded vowels.

This confusion in the perception of the basic vowel category has to be taken into account in the interpretation of the results, especially for the two rounded vowels, where it is strongest. Here, lip spreading for ‘smiling’ might interfere with the basic phonemic vowel quality. Another influencing factor could be the vowel intrinsic larynx height, which is e.g. in the case of /uː/ rather low compared to /aː/.

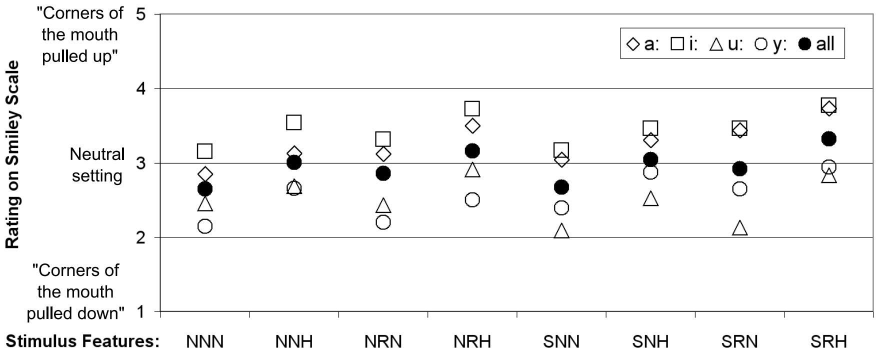


Figure 6.3: Mean values of the ‘smileyness’ of the four vowels. N = Neutral, S = Spread lips, R = Raised larynx, H = High f_0.

6.3.2 Detection of smiled vowels

The overall results of the formal perceptual evaluation are presented in Figure 6.3, which provides an overview of the mean values of the perceived degrees of ‘smileyness’ of the eight possible versions of each vowel. The corresponding mean values are listed in detail in the top half of Table 6.2, and the bottom half of Table 6.2 gives an overview of the average ratings sorted by activation or deactivation of the individual articulatory factors lips, larynx, and f_0. Results of the statistical analyses are briefly discussed below, while the details of the significance tests are listed in Appendix D.

For convenience, the features of the stimuli are coded in a shortened form as follows: [*lip feature*][*larynx feature*][*f_0 feature*]. The lips can be neutral or spread (N/S), the larynx neutral or raised (N/R), and f_0 can be neutral or high (N/H). ‘SRH’ indicates *s*pread lips, *r*aised larynx, and *h*igh f_0. See also Table 6.1 for an overview. If a whole group of stimuli is addressed, a dash indicates under-specification of the respective feature, e.g. ‘– – H’ denotes *all* stimuli with high f_0, regardless of the other articulatory parameters.

When we compare the ‘neutral’ baselines (NNN) of the four vowels with each other, /iː/ is perceived as the most ‘smiley-like’ and /yː/ as the least ‘smiley-like’. Overall, the stimuli with the highest scores, i.e. the highest degree of perceived ‘smileyness’, are those with *high f_0* (– – H). The best score is obtained by an activation of all three parameters (SRH), i.e. smiled lips, raised larynx, and increased f_0. This is true for all vowels in the data set except /uː/ where the combination with neutral lips scores best (NRH).

ANOVAs ($\alpha = 5\ \%$, details see Appendix D) reveal that overall the individual articulatory parameters of lips, larynx and f_0 have a significant influence on the rating. For some vowels, also the lip setting leads to significantly different ratings (*Vowel * lips* in Appendix Table D.1). Overall, the vowels are rated significantly differently, all except for /uː/ vs. /yː/ (Appendix Table D.2).

Table 6.2: Summary of the average ratings of the smiled vowels during the ‘smileyness’ rating task. Higher values indicate higher perceived degree of ‘smileyness’. **Top half:** Ratings of the individual stimuli. Odd numbers for *n* in /yː/$_{NRN}$ are due to missing values for 19 subjects because of technical problems during the test. **Bottom half:** Ratings averaged over individual articulatory factors (lips, larynx, and f$_0$). The factors are either neutral (*N*eutral) or active (*S*pread, *R*aised, *H*igh, respectively). Values in italics indicate opposite directions, i.e. when the activation of a ‘smileyness’ parameter induces a *lower* degree of perceived ‘smileyness’ (only in /uː/).

Individual	Feature	ID		/aː/			/iː/			/uː/			/yː/			Over all vowels	
stimulus	combinations		ave	sd	n	ave	sd	n	ave	sd	n	ave	sd	n	ave	sd	n
	All neutral	NNN	2.85	0.67	36	3.15	0.51	36	2.45	0.76	36	2.15	0.66	36	2.65	0.75	144
		NNH	3.13	0.83	36	3.54	0.59	36	2.69	0.57	36	2.66	0.86	36	3.00	0.80	144
		NRN	3.12	0.59	36	3.31	0.81	36	2.44	0.75	36	2.20	0.55	17	2.85	0.82	125
		NRH	3.50	0.60	36	3.72	0.76	36	2.91	0.70	36	2.50	0.88	36	3.16	0.88	144
		SNN	3.05	0.67	36	3.17	0.58	36	2.09	0.76	36	2.39	0.58	36	2.67	0.78	144
		SNH	3.31	0.71	36	3.46	0.62	36	2.53	0.85	36	2.87	0.89	36	3.04	0.85	144
		SRN	3.44	0.72	36	3.46	0.81	36	2.13	0.82	36	2.65	0.67	36	2.92	0.94	144
	All marked	SRH	3.73	0.67	36	3.77	0.85	36	2.83	1.03	36	2.94	0.78	36	3.32	0.94	144

Variable	Value	ID		/aː/			/iː/			/uː/			/yː/			Over all vowels	
(Factor)			ave	sd	n	ave	sd	n	ave	sd	n	ave	sd	n	ave	sd	n
Lips	Neutral	N– –	3.15	0.71	144	3.43	0.71	144	*2.62*	0.72	144	2.40	0.80	125	2.92	0.83	557
	Spread	S– –	3.38	0.73	144	3.47	0.75	144	*2.40*	0.91	144	2.71	0.76	144	2.99	0.91	576
Larynx	Neutral	–N–	3.08	0.73	144	3.33	0.60	144	2.44	0.77	144	2.52	0.80	144	2.84	0.82	576
	Raised	–R–	3.45	0.68	144	3.57	0.82	144	2.58	0.88	144	2.63	0.79	125	3.07	0.91	557
f$_0$	Neutral	– –N	3.11	0.69	144	3.27	0.70	144	2.28	0.78	144	2.37	0.65	125	2.77	0.83	557
	High	– –H	3.42	0.73	144	3.62	0.72	144	2.74	0.81	144	2.74	0.86	144	3.13	0.88	576

Regarding each vowel individually, the articulatory factor *spread lips* causes significant rating differences for /aː/ and /yː/. No differences are found for /iː/, as expected, because of the /iː/'s inherent setting of spread lips in its neutral form. Most notably however, /uː/ shows significantly *less* 'smileyness' with spread lips (2.62 vs. 2.40 on the rating scale, higher values indicate a higher degree of perceived 'smileyness', cf. values in italics in bottom half of Table 6.2).

The articulatory variable *raised larynx* causes a significant effect for /aː/ and /iː/ but not for /yː/ and /uː/. Here, for both rounded vowels a shortening of the vocal tract does not lead to an increased 'smileyness' perception. The factor *higher* f_0 causes significant differences on the perceptual 'smiley-scale' for all four vowels. In the individual vowel ratings, we find no interaction between any of the three articulatory variables.

6.4 Discussion

In this section, we discuss the empirical results, firstly addressing phonetic aspects of smiled speech (Section 6.4.1), and secondly discussing technical issues of the synthesizer related to this experiment (Section 6.4.2).

6.4.1 Phonetic aspects

When employing VocalTractLab to simulate fine vowel details, we found that 'smileyness' perception can be induced in all the vowels investigated in this experiment. Yet, the top-down articulatory strategy which has been pursued to imitate smiling needs further refinement. Our results suggest that it is not sufficient to exclusively manipulate lip spreading, larynx height (including voice quality), and f_0 in the way we presented here.

In the following, we address limitations regarding this experiment: f_0 manipulation, f_0 contour, interferences from unfamiliar sound qualities and the presentation of vowels in isolation, as well as interference of emotions with the technical idea of 'smileyness' on the articulatory level. Afterwards, we discuss aspects of the visual vs. the acoustic domain.

The overall significant effect of the f_0 parameter might indicate that our manipulation of neutral vs. raised f_0 was too coarse. We are trained in everyday communication to detect even small f_0 changes to extract the intonation contour from a speech signal. Thus, further experiments would perhaps benefit from using smaller f_0 manipulations, i.e. more intermediate values, to match the training and sensibility of the human ear for intonation. This would also better match the subtle character of the lip and larynx manipulations. A control condition could make sure that the participants did not 'learn' to assess f_0 instead of 'smileyness'.

Overall we found that *un*rounded vowels can reach higher 'smiley' scores than *rounded* vowels when activating all three 'smileyness' parameters. As an extreme result in this line we could regard the ratings for /uː/. In German, the vowel /uː/ is extremely rounded. When the feature *spread lips* was activated, the /uː/ stimuli always received *lower* scores than their counterparts with neutral lips, i.e. when larynx and f_0 parameters were kept constant.

It seems that the ‘contravention’ of roundedness weighs more than a possible signal of ‘smileyness’. This is perhaps due to the fact that the new vowel quality (close, back, *un*rounded) is too unfamiliar to German natives’ ears. Or, perhaps, our participants are interpreting this particular single vowel utterance primarily in emotional terms: Similar to the single vowel utterance “ah!”, which can express regret, appreciation or recognition, a back, close, unrounded vowel utterance often spelled <ugh> is associated exclusively with disgust. This was revealed by informal comments. For a natural smiled /uː/ we therefore hypothesize that humans do not use regular lip spreading to achieve perceived ‘smileyness’ but something else. It is possibly a combination of lip spreading and reduction of mouth aperture by pressing together the far ends of the lips on each side. This needs to be verified by analyzing video sequences with smiled utterances.

The issue with the smiled /uː/ may point to a general difficulty in the interpretation of the results, based on the fact that vowel quality is not very well preserved for some manipulated versions of /uː/ and /yː/. This might have led to confusions in the participants as to which vowel category they were listening to after all, to then be able to judge whether that assumed vowel was smiled or not. This is, of course, much less likely to arise with vowels embedded in an extended lexical structure.

The latter aspect is part of a more general limitation, namely that the perception of stationary vowels might not be directly comparable to the perception of fluent smiled speech. ‘Smileyness’ in fluent speech presumably does not always occur in extreme forms. The degree of ‘smileyness’ may depend on the changing levels of emotional state of the speaker, as well as on different sound categories: Some phones are possibly more exploited to convey ‘smileyness’ than others. Therefore, especially lip spreading without raising the fundamental frequency could be examined further, in a more natural stimulus environment such as whole words, to find out e.g. whether the low scores with /uː/ are found in *words* as well, or if this is an artifact of the isolated-vowel presentation mode. Dynamic changes *within* a vowel have already been shown to facilitate the perception of emotions (cf. e.g. Xu and Chuenwattanapranithi, 2007).

Participant feedback showed that the ‘smiley’ scale itself apparently invoked emotions in some of the subjects. They interpreted the upper end of the scale (5) as ‘friendly’ and commented: You could also be ‘friendly’ whilst speaking with “mouth corners pulled down.” This indicates a mismatch or interference of dimensions which has to be considered in future experiments. The invoked emotional interpretation interferes with the technical idea of ‘smileyness’ in terms of articulation.

A second f_0-related consideration addresses the overall level of the rating scores obtained from the stimuli, in conjunction with the mismatch of emotion vs. technical ‘smileyness’. Although significant differences in perceived degrees of ‘smileyness’ were obtained, the general mean values even for the ‘completely smiled’ settings (SRH) remain rather low. The fact that they stay below 4 out of 5 points on the scale could be to some extent due to the flat contour of the intonation in each vowel. As indicated in Section 6.2.1, the vowel stimuli were constructed with a monotonous f_0 contour, since in a pretest a more varied contour revealed associations with disgust, especially in the vowel /uː/, as discussed above. In general, this monotonous contour may have caused the low scores because smiled speech is not only associated with higher f_0 but also with more variability. Moreover, monotonous intonation contours sound artificial, and thus do not imply much of any form of positive emotion. Since a possible interference of emotion vs. technical 'smileyness' was found in the test, a monotonous intonation contour perhaps caused low scores also because no happiness, friendliness or politeness could be associated with the sound.

Aside from these limitations, it seems advisable to extend the set of smiled sounds, since it currently only comprises four vowels. This would generate the need for further manipulation parameters. For instance, Robson and MackenzieBeck (1999) observed a more "i-face"-like articulation for smiled open vowels, claiming that the vowels have a reduced jaw opening angle. To investigate the factor *jaw opening angle* with VocalTractLab, experiments with stimuli featuring reduced opening angles in different open and mid-open vowels could be designed. This would be a prerequisite to eventually generating a complete smiled vowel set in the articulatory synthesis approach. However, for consonants, especially those with labial activity such as /m p b v f/, the changes remain unknown and probably have to be studied separately.

Although this experiment is based on only four vowels, it seems to contribute further evidence to the following aspect of smiled speech. The main *visual* feature of human smiling is lip spreading (cf. Ekman and Friesen, 1978).[2] Yet, in the *auditory-acoustic* domain the visual importance of lip spreading seems to be different, i.e. smaller. Supporting Tartter (1980), Tartter and Brown (1994) and Robson and MackenzieBeck (1999), our results raise the assumption that the auditory characteristics of smiling during speech involve more than just a horizontal retraction of the lips, and sometimes even avoid lip spreading.

To sum up, extensions in future experiments should involve longer utterances than just stationary vowels, make use of parameters such as jaw opening angle (JA), and apply dynamic changes of parameters as suggested in Xu and Chuenwattanapranithi (2007). It is also advisable to integrate perceived phone quality into the perception test directly in order to use it as an additional variable for analysis.

6.4.2 Synthesis technical aspects

After the experiment in Chapter 5, where only larynx height was manipulated, the current experiment is a second piece of evidence indicating that VocalTractLab simulates fine articulatory detail on a very fine articulatory and acoustic level. The top-down applied 'smileyness' parameters all showed plausible acoustic consequences. However, a more suitable and robust articulatory manipulation strategy is needed to create 'smiling', since the manipulation

[2] This was found for both 'felt' and 'non-felt' smiles. However, we do not address the question here whether our participants perceived felt or non-felt smiles, cf. Ekman and Friesen (1978); Schröder et al. (1998); Drahota et al. (2008).

of only the three parameters lip spreading, larynx height, and f_0 obviously destroyed some phonemic sound categories, notably /uː/ and /yː/.

While technically, we found no limitations in the creation of our vowel stimuli, further experimenting might make one limitation obvious. As hypothesized above, ‘smileyness’ in the vowel /uː/ seems to, at least in parts, rely on pressing together the lips near the corners of the mouth. This, however, cannot be imitated with the current model of the synthesizer. This would presumably involve adding at least one more articulatory parameter to control the position of the lips.

Chapter 7

Experiment III – Vocal age in diphthongs

Even without seeing an unknown speaker, we automatically assess different personal characteristics from their speech. Besides gender, age is among the first assessments we make about an unknown speaker. We are able to do this merely based on hearing because, with aging, everyone goes through a number of physiological and hence acoustic changes of the speech production system, which manifest in the way our voice sounds. This is what we call the process of vocal aging. In this experiment this term addresses basic articulatory-phonetic features and excludes complex or higher-level parameters such as speaking rate or word choice. Within the framework of articulatory synthesis, vocal age is regarded as an interaction of different phonatory qualities with different supraglottal configurations.

We present a production and perception study of synthetic speech samples that are designed to represent male voices of three different age groups. Based on findings in the literature, several manipulation parameters are defined for YOUNG, ADULT and SENIOR voice age groups. Synthetic vowel stimuli (diphthongs) are generated, and evaluated in a perception test as to how well the intended age classes are identified by listeners.

As the two previous experiments indicated, VocalTractLab is capable of imitating subtle variations in the voice in a human-like manner. The control of the synthesis parameters and the quality of the acoustic output proved to be fine-grained enough to synthesize nuances of different degrees of breathiness linked to different vertical positions of the virtual larynx. Likewise, the smile-related changes in vocal tract shape induced appropriate nuances of formant shifts. While Experiment I and II each focused on few basic components in the voice, we now combine these manipulation dimensions, adapt them to the needs of the age-related synthesis task, and also incorporate new parameters such as jitter and shimmer, i.e. short-term perturbations of fundamental frequency and intensity. Thus, the final age-related stimuli comprise different combinations of f_0, voice quality, and settings of articulation that could be relevant for age modeling.

The speech production goal of this experiment is to suggest a parameter scheme for a simulation of three different age classes. The scheme is based on articulatory considerations and derived from literature on vocal aging. It is developed through exploratory feature analysis and immediate informal auditory evaluation, and afterwards evaluated acoustically and with a formal listening test.

The technical evaluation goal comprises a feasibility check of manipulating a complex set of vocal features, assessing the general usability of the manipulated parameters. Furthermore, we employ a special tool for batch synthesis of utterances with VTL. Its usefulness will be assessed.

In terms of the overview of all experiments (Figure 4.3), this experiment focuses on glottal and supraglottal parts of the speech apparatus, because we vary phonatory settings and parameters of the vocal tract shape.

In Section 7.1, we give a short background on findings regarding vocal age in human subjects, focusing on changes with age regarding articulatory and acoustic aspects of speech. On this basis, we derive the synthesis manipulation parameters. Section 7.2 describes the actual data creation and analysis procedures. The articulatory manipulation parameters are first tested to find suitable value ranges, which are then used for the generation of aged voice stimuli. The stimuli are subjected to acoustic and auditory evaluation procedures. The evaluation results are presented in Section 7.3, followed by a discussion of the phonetic and technical evaluation aspects of this experiment. We conclude by suggesting how the synthetic stimuli can be used within the framework of an automatic age classification system.

7.1 Vocal age in speech production

Findings in the literature show that vocal aging is a very diverse phenomenon, and results presented in one study may contradict those of another. This may be attributed to the increasing variability in voices as biological age increases, and the difficulty in separating normal aging effects from other factors such as lifestyle, fitness, psychological state or sociocultural environment (cf. Linville, 2001: 13). For instance, many aspects of speech-production related changes might not be caused by aging but by disease. These in turn might commonly be age-related but are, in essence, not part of the aging process itself. They are an example of *age cohort effects*, meaning that people born within a certain time share certain attributes. Thus, although in the following we describe changes that are *often* observed with aging, exceptions from this description are very likely to exist since age is easily confounded with other variables.

In the first section, age-related changes that are relevant to our experiment are described, followed by a description of how to transfer these age characteristics into settings of the synthesis software.

7.1.1 Human articulation and acoustics

Studies have reported a variety of characteristics of younger vs. older voices. Schötz (2007a) gives a concise overview of age-related changes (cf. also Linville, 2001). Major articulatory and acoustic changes with aging of the voice include changes in mean fundamental frequency, voice quality changes, supraglottal changes, and changes in the manner of breathing as well as speaking, each of which is discussed below.

Glottally one may find changes within the laryngeal structures which include ossification of the laryngeal cartilages, increased vocal fold stiffness, and reduced vocal fold closure. This may have effects on fundamental frequency and voice quality. Linville (2001) summarizes studies on changes in fundamental frequency indicating that, on average, the adult male voice is lower than the young male voice and also lower than the voice of elderly men. More specifically, Linville (2001: 172) reports a decrease in f_0 of about 10 Hz “from young adulthood to middle age”, and a rise of about 35 Hz towards “advanced old age.” f_0 seems to be a very powerful cue for perceiving vocal age (p. 199).

Regarding voice quality, the framework of the *hoarseness diagram* (Fröhlich et al., 1997) inspired the perspective of our voice quality considerations. The diagram incorporates measurements of the periodicity in a signal and the amount of additive noise, both of which are discussed in the following. Firstly, periodicity, or the lack thereof, can be characterized by features such as short-term variations in f_0 (jitter, frequency modulation) and short-term variations in amplitude (shimmer, amplitude modulation), as well as long-term variations such as standard deviations of f_0 and amplitude. Linville (2001) reports that the relation of short-term variations to aging have not yet been clearly established because of many interfering factors. The long-term variations, however, seem to increase with age. As a general tendency though, all levels presumably rise because the regulation and control of voice by the brain deteriorates with increasing age, thus the vocal fold vibrations become more unstable. Secondly, regarding additive noise, indications for rising levels of spectral noise in aging men are found which can be associated with incomplete closure of the vocal folds. This leads to increased levels of breathiness in older voices.

Supraglottal changes with age are caused by anatomical changes in the vocal tract shape. Among others, it is reported (Linville, 2001: 179ff) that a lowering of the larynx into the neck takes place, which causes vocal tract lengthening. This leads to an overall tendency of vocal tract resonance frequencies to be lowered.

Due to age-related changes in the respiratory system, where pulmonary function is reduced, the maximum intensity (amplitude) of isolated vowel productions seems to decrease with increasing age. A relation of age and mean intensity in *connected* speech could not yet be clearly established (Linville, 2001: 177).

Lastly, old voices may also show a different speaking style than younger voices (Schötz, 2007a, Linville, 2001). This might be caused by different hormone levels controlling brain activity, as well as sociolinguistic factors. Firstly, a general stiffness of the speech apparatus combined with lower activation levels in the brain often causes old voices to sound slower and slurred, mainly characterized by "articulatory imprecision" (Linville, 2001: 153). A centralized manner of articulation is observed, perhaps due to less control of the articulators. Secondly, the manner of speaking within one age group may be influenced by the life-long exposure to certain fashions of that age cohort. These influences all have their share in shaping the manner of speaking in terms of e.g. intonation, fundamental frequency, and word choice.

7.1.2 Transfer into synthesis features

So far only few studies have been conducted that deal with speech *synthesis* of aged voices, especially if modeling of voice quality or articulatory synthesis are involved (cf. Schötz, 2006: 138, for resynthesis see also e.g. Harnsberger et al., 2008, Schötz, 2003). Synthesis methods other than articulatory synthesis have been used to model e.g. 'old' voices, in particular formant synthesis (e.g. Schötz, 2006, Schötz, 2007b). They operate at the level of the acoustic signal, and allow much more direct control of the acoustic outcome than articulatory synthesis. There has also been work on synthesis based on articulatory parameters, e.g. regarding different voice qualities in general (Karlsson and Liljencrants, 1996). However, the parameters used are not directly available in VocalTractLab, such as formant synthesis parameters for direct control of the acoustic output; or detailed low-level articulatory parameters, such as an explicit modeling of the damping factor of the trachea (Karlsson and Liljencrants, 1996: 145). In other words, the details of how the voice simulations are controlled often do not match the articulatory categories of VTL to allow for a straightforward transfer. Thus, the following decisions on age parameters are a first attempt to simulate vocal age with VocalTractLab.

Motivated by the findings for human voices discussed above, we derive the following age cues for this experiment: Fundamental frequency (f_0), two voice quality (VQ) parameters, i.e. 'breathiness' (spectral noise) and 'roughness' (jitter, shimmer), and vocal tract lengthening. Their technical implementation details are described in the following. They are the results of extensive pretesting, which is discussed in Section 7.2.1.

The basic f_0 contour is created by providing a starting frequency value to the synthesis system. To generate a pleasant sounding contour, a declination component of 2 Hz per second is added, the decay component (see equation 7.3). The resulting contour is identical for each stimulus.

Voice quality is manipulated in two separate aspects (breathiness and roughness), which both represent articulatorily complex features. Firstly, 'breathiness' is determined by three articulatory factors: The basic distance of the vocal folds during phonation (degree of abduction, displacement), the width of the posterior chink or glottal leak, and the vertical phase difference of the vocal fold vibration (cf. glottis model in Figure 3.4). Vertical phase difference is added, based on above-mentioned considerations of stiffness of the vocal folds and because systematic testing of this parameter yields appropriate voice qualities.

Secondly, ‘roughness’ is determined by the two factors jitter and shimmer. Jitter is generated similarly to Klatt and Klatt (1990: 839). The authors propose to add a “slow quasi-random drift to the f_0 contour” by means of a control parameter named “flutter.” It mimics jitter but is not truly randomly organized:

$$\Delta f_0 = \frac{FL}{50}\frac{F0}{100}[\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t)]Hz, \tag{7.1}$$

where *FL* is the “flutter” component to vary the strength of the f_0 variation (Eq. (1) in Klatt and Klatt, 1990: 839). For better acoustic results regarding age related characteristics, and more freedom in variation, especially to simulate the older voices, we add high-frequency sine waves, alter the sine waves’ frequencies, and introduce varying offsets for them:

$$\begin{aligned}\Delta f_0 = {} & 1.4\sin(4 + 2\pi t) + JA\sin(3 + 2\pi 6.1t) + JA\sin(2 + 2\pi 13.1t) \\ & + JA\sin(5 + 2\pi 22.7t) + JA\sin(5 + 2\pi 1353.3t) + JA\sin(2.5 + 2\pi 2003.3t),\end{aligned} \tag{7.2}$$

where *JA* is the jitter amplitude to vary the strength of the f_0 variation for all but one sine wave component. The complete calculation of the intonation contour is based on the starting value for f_0, the decay component, always set to $f0_{Decay} = -2Hz/s$, and the jitter component:

$$f0_t = f0_{start} - f0_{Decay}\, t + \Delta f_0 \tag{7.3}$$

Regarding shimmer values, in basic tests we find that an explicit manipulation of the degree of *jitter* automatically influences the shimmer values to a considerable extent. To increase the shimmer values even more, however, we model shimmer in an analogous scheme to jitter. This influences the subglottal pressure curve and creates micro-perturbations in the intensity contour (amplitude modulation). Basic subglottal pressure is set to 800 Pa, upon which the shimmer component Δp_{sub} is added:

$$\begin{aligned}\Delta p_{sub} = {} & 15\sin(2 + 2\pi 4t) + 22\sin(3 + 2\pi 14.4t) + 30\sin(4 + 2\pi 21.3t) \\ & + SA_1\sin(5 + 2\pi 1005t) + SA_2\sin(2.4 + 2\pi 1378.3t),\end{aligned} \tag{7.4}$$

where SA_1 and SA_2 are flexible shimmer amplitudes to vary the level of subglottal pressure in two sine wave components.

Vocal tract length is manipulated by lowering the larynx with increasing age. For each basic vowel used in this experiment (/a ɪ ʊ ɔ/), we change the vertical position of the tongue root/hyoid bone parameter HY (cf. parameter overview in Figure 3.3b and Table 3.1), and sometimes adjust it horizontally in order to preserve vowel quality. Typically, the default vowel configuration of VTL is selected to represent the YOUNG setting. However, for /ɪ/, the larynx of the default setting is already nearly at the lowest possible point. Therefore, this is selected as the SENIOR setting and HY is increased for the other age classes. Auditorily, the obtained vowels sound different in paralinguistic quality, but not in phonemic terms. The higher-larynx samples sound ‘fresher’, the lower ones more ‘throaty’ and ‘tired’. Acoustic changes are plotted in Figure 7.1, p. 103 and discussed in Section 7.3.1.

We do not simulate breathing-related age changes nor changes in the manner of speaking. Changes of breathing, which e.g. lead to changes in loudness (intensity), are omitted in this experiment because in the speech technology environment loudness is a feature that is largely influenced by the distance of a speaker to the microphone, or the volume setting of loudspeakers. Additionally, the relation between aging and speech intensity is not clearly established yet (Linville, 2001: 177). Therefore, all stimuli are simulated with a constant sound pressure level. Secondly, we also omit variables regarding manner of speaking because the variations are too diverse to be simulated in this experiment and they lie outside the articulatory-phonetic goals of the thesis.

7.2 Data and analysis

For the perception test, we construct the stimuli in a very uniform manner, since regarding the perception of vocal age many ‘non-phonetic’ factors can influence age ratings, including factors related to the speech sample. Further factors are related to the task during the perception test, as well as speaker-related features and listener-related ones (see Schötz, 2006, for a comprehensive overview).

Kreiman et al. (2007: 2355) report that evaluation of pathological voices is “routinely” based on isolated vowels, e.g. samples of a duration of 1 second, extracted from the middle of sustained vowels. They “carry much information about the voice source” and their simple structure elicits “responses from listeners reflecting simpler perceptual strategies.” Therefore the results can be “more easily interpreted.” We adapt and slightly modify this concept and select diphthongs as the basic units for the listening test.

Each stimulus consists of two consecutive diphthongs, i.e. two out of the three German diphthongs /a͜ɪ a͜ʊ ɔ͜ɪ/. The pairings are /a͜ɪ a͜ʊ/ and /a͜ʊ ɔ͜ɪ/. A single diphthong has a duration of 1.3 s, when concatenating two diphthongs, we add 0.4 s of silence between them. The articulatory gestures that produce the diphthongs are the same for all stimuli. Thus, the time structure of the speech samples, resembling e.g. speaking rate, is always identical. Diphthongs are used because they are completely voiced, from beginning to end, and therefore carry a lot of information about the voice source, and thus, possibly, age. We favor diphthongs over steady-state vowels because diphthongs also carry *time-dynamic* information on articulation, which makes listening to them more diversified.

The age groups are clustered in the following way:

YOUNG males: 15-24 years – ADULT males: 25-54 years – SENIOR males: 55-80 years

The age-class scheme is not primarily motivated by articulatory aspects. This would have yielded different age boundaries since relative *sharp* cuts take place at the age of puberty and when entering really advanced age. Instead, this scheme is based on a scheme used in automatic age classification, introduced by Müller (2005) and later applied in other studies on automatic speaker age recognition (see also Section 7.4.3). Using this technology-oriented scheme makes the stimuli compatible with the automatic age classifier described in Müller (2005), so that they can be evaluated by an automatic age classification task as well as a human listening task. In this thesis, we focus on the human listening evaluation but the discussion section briefly deals with a setup for automatic age classification (Section 7.4.3).

The set of age-related synthesis features introduced in the previous section has not been tested systematically before. Thus, we first determine the possible ranges of values for each feature that produce audible acoustic differences without introducing synthesis artifacts like inappropriate friction noises (Section 7.2.1). Within these ranges of values, we then define *sub-ranges*, one for each age class, to create the stimuli for the listener tests. The details of the age class settings and their corresponding stimulus creation are described in Section 7.2.2. Details on the evaluation setup are put forward in Sections 7.2.3 and 7.2.4.

We use the VTL GUI (see Section 3.1) as well as a batch tool which connects to VTL by an application programming interface (API, cf. Section 4.2.4) to explore and determine the exact age cue settings, and later to synthesize the stimuli for the perceptual evaluation. The batch tool has the advantage that a large number of systematically varied stimuli can be generated automatically for auditory evaluation. This way a relatively fast and convenient manner of exploring the feature space is possible.

7.2.1 First generation of stimuli: Feature exploration

The equations presented in Section 7.1.2 represent the result of extensive testing and auditory evaluation of different voice characteristics. The basic structure of the equations as well as their details, such as sine wave offsets and frequencies, were not clear from the beginning. Therefore, we initially developed varying definitions of the age-related parameters, and explored their feature spaces. In total, we generated and informally evaluated about 1200 different voice quality profiles using the batch tool. This yielded a preliminary set of stimuli. We briefly describe some landmarks of creating that set of stimuli, although it was discarded in favor of a second set that was found to be much more adequate in modeling age impressions. We point out the lessons learned, while the next section presents the final set of stimuli.

Overall, the voice qualities of the preliminary set of stimuli seemed acceptable, since pre-test listeners gave feedback that they could indeed hear differences in the voices. It was also reassuring to hear from pre-test participants that some of the voices indeed sounded like people they knew, when talking on the phone with them. However, pre-test listeners also commented “this sounds like a sick young adult” on stimuli that were meant to represent a SENIOR voice. Thus, we obviously had a problem of confounding fitness with age.

We therefore went through the modeling process again, weighing the individual age cues differently than before (discussed in Section 7.2.2).

Compared to the final stimulus features (cf. Table 7.1), the parameters in the preliminary set covered a much wider range, i.e. the different values assumed for a feature would exhaust a large part of the parameter space provided by VTL. Each parameter, applied in 'isolation', sounded appropriate in terms of age-related changes in voice quality. However, *combining* the parameters in one stimulus caused 'extreme' listening impressions and produced unexpected interactions, e.g. introducing synthesis artifacts such as friction noises. For instance, a large displacement of the vocal folds used for the SENIOR voices in combination with certain vowel-vowel transition durations would, to a large degree, amplify synthesis artifacts. Other overemphasized factors included large vertical phase differences, resulting in impressions of diplophonia, and the size of the posterior chink (glottal leak).

Since the basic units of the stimuli were diphthongs, we tried producing them at a rather slow articulation rate, so they would reach a relatively long overall duration. This was intended to give the listeners more time to gather an impression of the voice that they had to classify. As indicated above however, the voice quality settings interacted audibly with the supraglottal vowel gestures. Therefore, the gestural transition duration from the first to the second vowel of a diphthong had to be carefully adjusted because too long transition phases introduced prominent friction noises and similar other noisy artifacts.

As part of the learning-by-exploration process, we reduced the range of the feature space for each age class, and carefully adjusted the vowel-to-vowel transitions in the diphthongs, to produce a more acceptable data set for the final age classification test. A theoretical motivation for reducing the feature space was the idea of aiming at representative age tokens that were situated at the center rather than at the edge of a particular class. In other words, we aimed at imitating typical young, middle-aged and old sounding voices rather than atypical speakers which are 'young but very adult sounding', 'middle-aged but rather old sounding', or 'youthful sounding Seniors' (regarding typical and atypical speakers, see also Schötz, 2001, 2004). It is hard to reliably judge, though, how close a feature value is located to an age-class boundary because of a lack of experience with this kind of data.

7.2.2 Creation of aged voice stimuli

In the following we describe the creation of the final stimulus set which simulates the three different age groups YOUNG, ADULT, and SENIOR with the age class boundaries given on page 96. To avoid making the parameter changes too obvious for the ear, we generate more than one representative voice quality for each age class, applying a 3-factorial design (3 x 2 x 2 = 12), which specifies 2 or 3 levels for each of the following age cues in each age class and produces 12 voice qualities per age class:

Table 7.1: Values of the age-related synthesis parameters per age class. Details regarding glottal parameters, such as parameter ranges, can be found in Table 3.2, p. 33. Jitter amplitude JA and shimmer amplitudes SA_1, SA_2 are introduced in equations 7.2 and 7.4.

Cue	Level	Parameter	Age Class Values		
			Young	**Adult**	**Senior**
Pitch	High	f_0 [Hz]	130	115	150
	Middle		120	107.5	127.5
	Low		110	100	105
VQ: Breathiness	Breathy	Displacement [mm]	0.4	0.6	0.8
		Glottal Leak [mm^2]	-1	1	2
		Vertical Phase Lag [rad]	0.1	0.2	0.3
	Modal	Displacement [mm]	0.5	0.7	0.9
		Glottal Leak [mm^2]	0	1.5	2.5
		Vertical Phase Lag [rad]	0.2	0.3	0.4
VQ Roughness: Jitter	Regular	JA	0.3	0.5	1.3
	Irregular		0.4	0.6	1.6
VQ Roughness: Shimmer		SA1, SA2	30, 40	40, 50	50, 60
Larynx		Vertical Position	High	Medium	Low

- 3 levels of f_0: High (H), Middle (M), and Low (L)
- 2 levels of VQ ‘Breathiness’: Modal (m) and Breathy (b)
- 2 levels of VQ ‘Roughness’: Regular (r) and Irregular (i)

Table 7.1 gives an overview of the particular settings in each age group. The different values for each class indicate that each age group is represented by 12 slightly differing *voice profiles*, which are completely distinct from the voice profiles of the other two age classes. The distinctiveness, or idiosyncrasy within the age groups means that the parameters do not undergo a true permutation of all values across all age groups. The goal was to create *typical* voices rather than to permute all combinations.

The continuum of reasonable sounding feature values is based on the literature, on acoustic measurements and on subjective listening impressions. The leading hypothesis for our simulation of synthesized speaker age is that the ‘older’ a voice is intended to be, the more ‘marked’ are most of its production parameters (see e.g. Kreul and Hecker, 1971). For our features this means that with advancing age (age class), we use increasing breathiness and roughness in voice quality, and increasingly lower the larynx. The particular levels of VQ ‘breathiness’ and jitter are derived from pretesting as indicated in Section 7.2.1. Additionally, each class features an age-class specific set of shimmer amplitudes and an individual vocal tract shape configuration for the vowels to introduce different supraglottal articulation settings for each age group. The resulting shimmer values in the speech signals of each age class are close to the values reported in Müller (2005).

The levels of f_0 for an age class are distributed evenly within the range of f_0 values typical of that age class (following Linville, 2001), resulting in a High (H), Middle (M), and Low (L) level for each class. Moreover, since f_0 is a leading cue for listeners, we pay particular attention to letting the values for the age groups overlap at least partly, so that human listeners should not be able to rely on this feature alone to make their age judgments.

Building upon the three-factorial design and creating 12 voice profiles for each age class, we thus obtain 36 voice profiles in total. Each of the 36 age profiles is represented by 2 different pairings of the diphthongs, yielding 72 different test stimuli in total. The pairings are /a͜ɪ a͜ʊ/ and /a͜ʊ ɔ͜ɪ/.

All stimuli are post-processed with Praat (Boersma, 2001) in the following way: They are scaled to 70 dB intensity, Hann-window pass-band filtered resembling telephone settings (0.3 to 3.4 kHz, 100 Hz smoothing), and resampled from 22050 Hz to 11025 Hz. Applying this kind of filter presumably increases the perceived naturalness of the synthetic speech because the degradation of signal quality appeals to the imagination of the human listeners. When listening to a slightly degraded signal, the missing acoustic components are filled in by the listeners themselves.[1]

7.2.3 Acoustic analyses

The acoustic analyses comprise a traditional formant analysis and an extensive analysis of different voice quality features. Since the vocal tract shape is slightly different for each age class, we analyze the formant frequencies of the vowels to provide information on the effects of supraglottal manipulation. We measure the formant values at the beginning of a diphthong and at the end, thus measuring the start and the end vowel of each diphthong. Using Praat, we extract the mean values of about 30 ms duration each, at a point where the vowel is fully audible but stationary, avoiding the transition phase between the vowels. We use standard formant analysis settings, only adapted to male voices, i.e. searching for 5 formants within a frequency window with an upper bound of 5 kHz. Additionally, we extract a mean f_0 value during each of these 30 ms intervals, using standard settings.

The analysis of voice quality features is executed using Praat's *voice report* function, with standard settings. The features include f_0 measurements, jitter and shimmer values, and harmonic-to-noise ratios (mean HNR). The analysis is performed on two data sets. The first analysis is done on the set of diphthongs as they are created with VocalTractLab. The values are provided to illustrate the acoustic nature of the voice profiles and thus the direct effects of the age-related manipulation parameters in VocalTractLab, with no external manipulation applied. A second analysis is performed on the post-processed diphthongs, which have undergone telephone filtering and downsampling. The values are presented to characterize the acoustic nature of the stimuli as they are presented to the participants of the listening test.

[1] The telephone filtering also increases the similarity of our stimulus material to the material that is used to train the age classifier described in Müller (2005), i.e. conversational *telephone* speech.

7.2.4 Listener test design

To obtain a perceptual evaluation of the simulated aged voices, we conduct a web-based listening test presenting a three-class decision task with the possibility to listen to a stimulus as often as a participant wants. The categories for classification are labeled as "YOUNG (15-24 yrs)", "ADULT (25-54 yrs)", and "SENIOR (55-80 yrs)." In each trial, we present one stimulus, i.e. a concatenated pair of diphthongs as described in Section 7.2.2, and the participant has to make a choice between the three age classes. The framework used is the web-based interface WebExp (Keller et al., 2009, cf. also Section 4.2.7). The layout of the test slides is similar to the one portrayed in Experiment II, Chapter 6. Instead of a rating scale, the participants see the three class labels and have to input one out of three letters, one for each age class (Y, A, S).

28 invited, German native participants took part in the test (10 female, 18 male, average age 37 years, SD 13 years, median 32 years, range between 24 and 67). Since the age of the participants themselves can influence their age judgments, we report their ages in more detail here: 1 participant belonged to age class YOUNG (below 25 years), 22 participants belonged to the age class ADULT, and 5 listeners belonged to the SENIOR class (above 54 years).

At first, the participants were familiarized with the technical procedure and the range of different stimuli by going through an explicit warm-up phase. We did not save any answers here nor did we provide hints on how to give 'correct' answers (cf. also Section 4.2.7). In the main phase of the test, each stimulus was judged twice, yielding a total of 144 stimulus judgments. To minimize order effects, all stimuli were randomized within blocks for each participant. At the end of the test, listeners could optionally give feedback on issues including perceived naturalness and perceived pathologies of the voices, and on the test procedure in general. We checked the results for validity and cross-speaker consistency. No participants were excluded from the analysis. The results are statistically analyzed using the chi square test.

7.3 Results

First, we present the results of the acoustic analyses of the 'aged' stimuli (Section 7.3.1), followed by the results of the age classification test (Section 7.3.2).

7.3.1 Acoustics

We describe the main characteristics of the acoustic properties across age classes as well as across processing conditions, because the telephone-filtered values differ from the ones derived from the non-filtered diphthong utterances. For reference, details can be found in Appendix E (Tables E.1 to E.3 for the individual voice profiles as they are created with VocalTractLab; Tables E.4 to E.6 *after* the telephone filter has been applied).

Across age classes, as intended, the general tendency in all measured categories is that the values increase with increasing age (age class), except for mean f_0 values, which, as desired, overlap across the age classes. HNR values decrease with age, indicating an increase in noise with older voices. HNR differences within individual voice profiles are mainly due to the different vowel qualities in the stimuli. The articulatory manipulations of voice quality thus seem to be reflected in the acoustic output.

Across processing conditions, telephone filtering and downsampling to 11025 Hz yields a considerable increase in shimmer, slightly increased values of jitter, and somewhat lowered harmonicity compared to the unprocessed version of the diphthongs.

Figure 7.1 illustrates the distribution of the vowels in the F_1-F_2 formant plane. For reference, Tables E.7 to E.12 in Appendix E list the corresponding mean formant frequencies, complemented by F_3 and f_0 values for each basic vowel /a ɪ ɔ ʊ/ of each diphthong /a͜ɪ a͜ʊ ɔ͜ɪ/ in each voice profile. The formant plot generally reflects the articulatory manipulations regarding larynx height. Please recall that the supraglottal manipulation feature *larynx height* (HY) is spread over the age classes in the following way: From YOUNG via ADULT to SENIOR, the larynx is lowered stepwise to the lowest position possible in VocalTractLab. The acoustic effects can most clearly be seen for /a/, where SENIOR age class vowels have lowest values in F_1 and F_2, which represents the acoustic correlate to lowered larynx position (the longer the tube, the lower the formants). ADULT age class vowels are in medium position, both acoustically and articulatorily, and YOUNG age class vowels are at the higher end. These differences are less pronounced for /ɔ/ and /ʊ/. For /ɪ/ the differences are even smaller.

The acoustic results can be explained by the details of the manipulated articulation: Variation of vertical larynx position (HY) is as follows: $HY_{Neutral} = -4.75$, $HY_{Highest} = -3.50$, $HY_{Lowest} = -6.00$. See also the overview of VTL parameters, Table 3.1, p. 29. In the case of /a/, the default (and YOUNG) vowel configuration in VocalTractLab has a relatively high larynx position, the highest of all four vowels under consideration ($HY_{\mathrm{a}} = -4.37$). Therefore, the larynx can be lowered quite a long distance and thus causes substantial acoustic changes. All other vowels show a lower YOUNG larynx position ($HY_I = -4.83$, $HY_U = -4.87$, $HY_O = -5.20$), and some show a very narrow pharyngeal cavity (/ʊ ɔ a/. Narrowing of the back cavity lowers F_2 (cf. Neppert, 1999: 133), thus /ʊ/ and /ɔ/ have smaller capacities for F_2 decreases since the default values are already low. /ɪ/ has a relatively wide pharyngeal cavity, which already yields a relatively low F_1.

7.3.2 Listening test classification task

The answers of the 28 participants are shown in Figure 7.2 and used to generate a confusion matrix (see Table 7.2). The rows correspond to the intended age class, while the columns correspond to the perceived class (listener judgments). Hence, the diagonal (bold numbers) shows the percentages of correctly classified samples. Overall, the answers are highly sig-

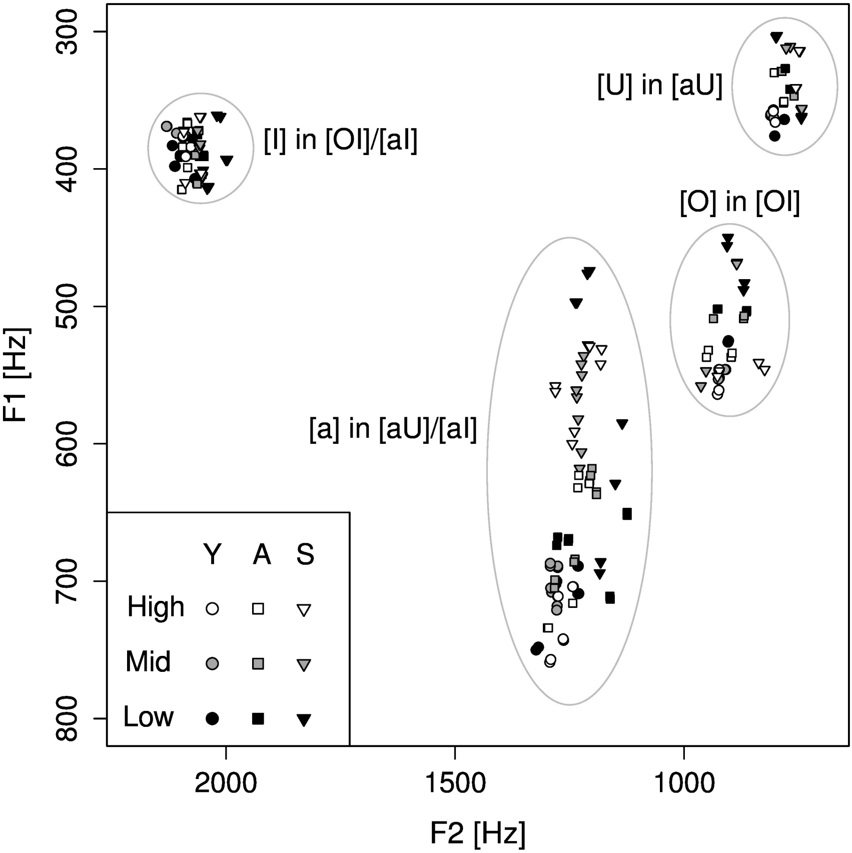


Figure 7.1: F_1-F_2 plot of the target vowels of the diphthongs in the age-related voice profiles. The labels indicate from which diphthongs a given target vowel originates (SAMPA notation). The shape of the ellipses is not based on standard deviations, but only to point out the different vowel regions. Different markers are used to indicate the different age classes, different shades are used to indicate the different levels of f_0.

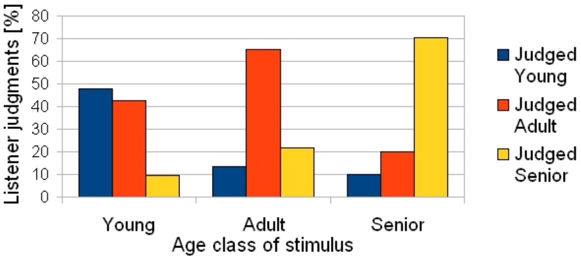


Figure 7.2: Listener judgments for each age class, based on 4032 judgments from 28 participants. Cf. also Table 7.2.

Table 7.2: Confusion matrices illustrating the age decisions of the 28 participants of the forced-choice classification perception test. Number of votes per age class (top) and distribution of votes in percent (bottom). Bold face indicates correct judgments per age class. The numbers are based on 4032 votes. The chance level is at 33.33 % since the subjects could choose between three age classes. Overall, **61.11 %** of the samples were classified correctly.

Age class	Judged Young	Judged Adult	Judged Senior	Total
Young	**644**	574	126	1344
Adult	179	**874**	291	1344
Senior	132	266	**946**	1344
Young	**47.92 %**	42.71 %	9.38 %	100 %
Adult	13.32 %	**65.03 %**	21.65 %	100 %
Senior	9.82 %	19.79 %	**70.39 %**	100 %

nificant ($\chi^2 = 1654.82$, $p < 0.001$), with an overall classification accuracy of around 61 %. Assuming that the age classes are situated on an age *scale* and having manipulated the features in principle by increasing their values with age, we assume an ordinal scaling of the classes. Then the results show a Spearman correlation coefficient of 0.544 ($p < 0.001$) between samples and judgments.

The numbers in the confusion matrix (Table 7.2) indicate that the Young samples are judged correctly in about 48 % of all cases. Roughly 43 % are misclassified as Adult, and roughly 10 % as Senior. The Young samples with non-high f_0 are often misjudged as Adult. Simulated Adult voices are mostly judged correctly as Adult. The samples that are most consistently judged correctly are the Senior samples. For reference, a detailed illustration of the voice confusion patterns can be found in Appendix E (Figure E.1, Table E.13). It shows the listener judgments in detail for each voice profile.

Further analyses of the perception patterns indicate which synthesized age cues apparently constitute the most representative voice profiles. The question is whether the voice profiles that receive the highest amount of correct age class judgments (top 25 %, i.e. top 7 samples) show a homogeneous pattern regarding the age-related features. The age decision

Young is most distinct for stimuli with high Young f_0 setting and 6 out of the 7 samples are worded /aɪ aʊ/. The top candidates for Adult decisions show no homogeneous pattern. The group of top Senior samples is dominated by the low Senior f_0 setting and also /aɪ aʊ/ wording, as in the top Young samples.

Since feedback from the participants in our study suggested that they would have appreciated more time to adjust to the synthetic nature of the voices, the data are also analyzed considering temporal aspects of presentation. As mentioned above, once all the 72 stimuli had been presented during the test, they were presented a second time. Between the first and the second half of the test, the accuracy increased from 58 % to 64 %. The corresponding confusion matrices and graphs are provided in Appendix Table E.14, Figures E.2a and E.2b. The increase in correct answers may indicate that basic features in the artificial voices do indeed portray characteristics that people associate with vocal age or at least synthesized vocal age.

In summary, the results suggest that our initial scheme of age-related synthesis features, based on literature findings, is, in general, ‘identifiable’ by human listeners. First inspections suggest that, in our data, f_0 seems to be a strong cue for synthetic age, as it is in natural (human) speech. Reports from participants also suggest that high jitter and shimmer values are a cue for the Senior age class.

7.4 Discussion

We discuss the empirical results firstly by addressing phonetic aspects of vocal aging (Section 7.4.1), and secondly by discussing technical capabilities of the synthesizer including the batch tool used (Section 7.4.2). Finally, we discuss how the stimuli can be used with an automatic age classifier (Section 7.4.3).

7.4.1 Phonetic aspects

In this experiment, we suggested a set of parameters regarding voice quality, fundamental frequency, and vocal tract shape to imitate possible vocal characteristics of three age classes (Young, Adult, Senior). Regarding the speech production goal, we found that our listeners were successful in identifying the correct age class, suggesting a general appropriateness of the age-related simulation scheme.

Phonetic details

Regarding the relative importance of individual cues for human perception of synthetic speaker age, our analyses indicate that f_0 was a strong cue towards the age decisions. f_0 is reported as being a “very powerful and resilient cue” (Linville, 2001: 199) in the perceived age of natural speech, too. The wording of the stimuli (/a͜ɪ a͜ʊ/ vs. /a͜ʊ ɔ͜ɪ/) seems to have influenced the decisions to some degree as well.[2] In addition, participants reported that 'rough' voice quality (high values of jitter and shimmer) influenced them to vote for the Senior age class. In combination with the high consistency in rating the Senior stimuli, it can be argued that the current way of imitating vocal age creates easily identifiable Senior voices. Young stimuli, however, were often recognized as Adult.

The decision pattern in favor of the Adult class may have to do with the age class boundaries. The age classes used in this experiment do not capture very well some critical anatomical shifts occurring with age, where substantial changes may occur around 40 to 45. So many of the more 'distinct' age differences reveal themselves only for the Senior age group, while the Adult group is very inhomogeneous. Determining the age boundaries for the classification task is difficult. Even if one follows anatomical developments with age, the picture is not clear cut. A solution to (avoid) this problem could be achieved by performing an age *estimation* task rather than an age class decision task.[3]

However, Adult answers could also have been essentially 'negative answers', meaning the stimulus sounds 'neither really old nor sufficiently young'. This could explain why the acoustic cues for the top-rated Adult stimuli do not show a homogeneous pattern. Furthermore, the frequent Adult answers might be an effect of the age class sizes. The age class intervals in terms of years spanned by each class is not evenly distributed, especially with regard to the Young class: The Young class spans 10 years (15–24), the Adult class 30 years (25–54), and the Senior class spans 26 years (55–80). The Adult class perhaps spans a too wide range of years. Thus, a large *variety* of voices is actually subsumed under this one class label because much can happen in 30 years. Especially since around the age of 40 anatomical changes can be large.

An effect of listener's age on their age ratings, as discussed in Linville (2001), seems minimal, since most of the listeners belonged to one age group (Adult: 22 listeners, Senior: 5, Young: 1). Linville (2001: 192f) reports that the rating performances seem to be "adversely affected by advanced age" and also by young age, perhaps due to, among other things, hearing loss and lack of listening experience. Another small influence may be the fact that the default voice of VTL is that of an Adult male speaker. This may attract Adult votes, but it did only do so with Young stimuli. The Senior voice profiles only rarely received Adult votings.

The design of our stimulus set does not enable a deeper analysis to find out which features had the greatest impact on the age decisions. The reason is that the ranges of the parameters are age-group specific for voice quality 'breathiness' and 'roughness' as well as

[2] Some participants reported that the choice of diphthongs evoked associations with age, in particular /a͜ɪ a͜ʊ/ evoked a picture of pain and weakness, which made them submit a Senior vote, according to self-assessments. However, when asking verbally, after a rating task, which features invoked particular decisions, the stated features "do not always agree with acoustic measures found to actually correlate with perceived age estimates." (Linville, 2001: 190)

[3] Perception tests to assess the perceived age of a listener can have the shape of age class decision tasks with e.g. two or three age categories (*Young–Old* or *Young–Middle Age–Old*), or they can be age estimation tasks, in which participants are asked to indicate the perceived age of the speaker in years. The more fine-grained statements the participants have to make, the more difficult they find the task (cf. Linville, 2001: 191, see also Schötz, 2007a).

vocal tract shape. Only for f_0 do values overlap, but they are still not identical for each age class. Thus no parameter is truly permuted across all age classes. Therefore, the parameters can only serve as a basis to find out whether these settings indeed enable listeners *overall* to make correct age class judgments. Future studies could adapt the proposed scheme to find out about the *individual* influences of each of the manipulation parameters, also including control conditions.

Regarding naturalness, the combination of the artificial voices with the telephone filter post-processing method seems to be a good choice. The perceived naturalness seemed to be of such a high degree that listeners reported that some voices sounded surprisingly similar to people they know. It was also reported that especially older voices seemed to sound very authentically aged.

The increase in classification accuracy from the first to the second half of the experiment indicates a learning effect, a common psycho-acoustic phenomenon having to do with experience (see also Section 4.2.7, p. 60). Since the listening test participants had never heard these artificially aged voices before, they apparently needed some time to get used to the artificial vocal age continuum. On the one hand, it is not clear what cues the listeners responded to most strongly. On the other hand, the increase in correct answers indicates that basic features in the artificial voices do indeed capture characteristics that people associate with vocal age. The learning curve or adjustment speed is also influenced by stimulus length and variability. Linville (2001: 191) reports that the percentage of correctly classified stimuli depends on, among others, how long the samples are and what kind of speech material is presented. If we used longer and more variable stimuli, accuracy would presumably rise even more.

Relation to other studies

Is is difficult to rule out confounds of age with other factors (cf. e.g. Schötz, 2006, Linville, 2001). So the results of many studies, including this one, should be regarded with some caution because it may not be clear what exactly they have been measuring. The variable *age* can easily be confounded with other variables like *fitness*, *sociolect* or *emotion*. This can be connected with the issue of modeling individual rather than age differences (cf. Schötz, 2006: 165). Since some studies concentrate on research of one voice per age class, the question arises which factors are age specific and which are idiosyncratic and speaker-specific. In our study, we tested voice quality for male sounding voices, all based on the same vocal tract anatomy and uttering the same speech material. Therefore, individual factors should be at a minimum level and the perceived age differences can thus be attributed to the changes in the manipulated parameters.

We can relate our overall accuracy results to outcomes of studies where perceived age was tested against chronological age based on natural speech data, as e.g. reported by Linville (2001: 191). In comparable tasks, an overall accuracy of 51 % is reported for phonated vowels, classified into the three categories *Young/Middle-Aged/Old*. The accuracy increases when only two categories are available (*Young/Old*, 78 %) and when read speech instead of phonated vowels is presented (87 % and 99 % in the two-category tasks). In our test, an overall accuracy of 61 % was reached. This relatively high result for phonated vowels can in large parts be attributed to the correct judgments of the SENIOR age class stimuli.

Listening test procedure issues

Finally, we need to address a technical issue in the listener result data: Automatic playback was not working equally well on every listener's computer, and participants reported that they sometimes accidentally clicked on an age class button twice which means that they rated two subsequent stimuli without having heard the second one of the two. This was probably due to failure of the automatic playback function. These incidents might have skewed the data to some extent. To characterize the impact, we counted the amount of ratings which were submitted after a presentation time of below 3.0 seconds, i.e. the duration of the actual stimuli. Ratings submitted before the obvious end of stimulus playback sum up to 13. Another 5 stimulus slides show a presentation time between 3.0 s and 3.2 s. Assuming the additional 200 ms as minimal reaction time required to plan and execute a mouse click or key stroke, 18 out of 4032 ratings (0.004 %) were submitted prematurely, presumably due to technical problems.

7.4.2 Synthesis technical aspects

With respect to the synthesis system's evaluation (technical goal) we find that we were able to manipulate a complex set of vocal features with VTL to produce complex arrangements of vocal qualities. This can be seen as an extension to the findings in Experiments I and II where less complex vocal settings were found to be manageable with VTL.

During the manipulation of the age features, we employed the *articulatory* synthesis system in a 'hybrid' way, by using both articulatory parameters and signal-processing methods. Some of our features were manipulated *articulatorily* (e.g. glottal leak, vocal fold distance, vertical phase difference, larynx lowering), leading to acoustic changes mainly in breathiness (HNR) and formant structure. Since the roughness values had changed only minimally, we added e.g. jitter by directly manipulating the f_0 contour, which is a signal-processing procedure, operating on the acoustic level, rather than an articulatory synthesis method.

The manipulated parameters were in general usable regarding simulated speaker age. However, the more deviating a setting was from the default voice, the more susceptible the output was to synthesis artifacts. Furthermore, the management of the parameters was in practice only made possible because we were able to use a special tool for batch synthesis of utterances. It considerably facilitated the systematic exploration of phonatory parameters by easily creating systematic ranges of voice qualities. Listening evaluation then led to the selection of the desired qualities. The tool proved to be useful in complementing the control of VTL through the graphical user interface (GUI). However, the tool is currently restricted to the particular set of manipulation parameters and diphthong utterances used in this experiment.

7.4.3 Automatic age classification

If the long-term aim is to create more natural and more individual sounding voices with articulatory speech synthesis, age is certainly an aspect one should be able to simulate. Thus, investigating vocal age in speech production can be beneficial not only for speech research but also for the area of speech technology. The findings gained from these kinds of experiments can be useful to improve speech synthesis and speech recognition systems (cf. Bocklet et al., 2008), systems for security applications (e.g. voice identification) or medical applications, e.g. voice diagnosis tools that remotely distinguish between a normally aging voice and a voice that has become pathological.

Furthermore, knowledge about age-related voice features can improve automatic age classification systems (but cf. also Wolters et al., 2009, on the lack of reliability of acoustic features). Based on their classifications, dialog systems can automatically adjust the dialog style to the user's age. Thus, although the experiment presented in this chapter is phonetically motivated, it can be helpful in the area of speech technology because the phonetically and theoretically motivated nature of the stimulus set can support error analyses of automatic age classification systems (Doddington et al., 1998, Shriberg, 2007).

In our case, we can use the precisely controlled stimuli to evaluate a classification tool which produces individual age class scores. To obtain high degrees of classification accuracy, the classifier's age model parameters are tuned on large amounts of training data. The resulting models work well on similar evaluation data but it is not clear exactly what kind of age characteristics they picked up from the training data. By running automatic evaluations with the stimuli of this experiment, one could look into the black-box-like architecture of the classifier's age models so that they can be further improved (cf. also Feld, 2011: 149ff). Since we know the system's performance on natural human data, by analyzing the patterns in the recognition score output, we may find insights that help in improving the recognition system in a way not possible before. For further refinement, one could run age classification experiments with control conditions, i.e. manipulate only one parameter while keeping the others constant. To do this with human listeners would be very time consuming.

Since the classifier is trained on natural human data, its age-related rating scores could also support the development of more naturalistic synthesis stimuli. But this improvement could also easily lead to stimuli which are over-adapted to the classifier.

In the current experiment, we used our stimulus set of 36 voice profiles in three age classes as input to an automatic age classification system, described in Feld (2011) (see also Müller, 2005). Since the classifier has been trained on seven age models (3 male-voice age classes, 3 female, 1 child), it has to decide between 7 classes, whereas the human listening task only represented a 3-class decision problem. The results of this automatic classification are presented in Table E.15 and Figure E.3 in Appendix E. The recognition scores are low, but about twice as high as chance level.

We would like to point out two aspects of these results. Firstly, the system-inherent age models were originally trained on conversational telephone speech, i.e. *fluent human* speech, and they were deliberately *not* tuned to the synthetic voices or the diphthong word material in order to both prevent overfitting and to allow for a more authentic insight into the quality of the synthetic voices compared to natural voices. This is presumably the main reason for the overall recognition scores being low. In this setting, however, the recognition results indicate a strong influence of f_0 and of the wording – /a͜ɪ a͜ʊ/ sounds consistently older, /a͜ʊ ɔ͜ɪ/ younger – as has been noted for the results of the human listening test (Section 7.3.2, p. 105).

Secondly, the classification data indicate another match with the human results, namely regarding the observation that the ADULT class seems to be a container class for stimuli that do not fit YOUNG or SENIOR class: When sorting the ADULT samples as a function of score, no particular voice quality pattern representative of ADULT voices can be determined. This is a parallel to the finding that in human ratings, the top-rated ADULT samples show no homogeneous pattern regarding the parameter values.

Although the initial classification scores are low, the basic setup can be used for further evaluation of the automatic classifier: High-level articulatory features could provide deeper insights into the age models and enable improvements of the original low-level classifier features. However, to achieve this, our stimulus set would have to be extended and would have to incorporate still more systematic variations in the age classes.

To conclude, we point to a fundamental difference between the phonetic and the speech technology approaches. While for the speech sciences it is important that ‘age’ is extracted in a maximally clean way from anatomy, articulation and acoustic speech data in order to *generalize* the findings, speech technology rather aims to find cues that determine which age class people most likely belong to. This can be regarded as a contrast between searching for age factors per se (and not age cohort effects) vs. aiming to develop a working system where the time frame is *here and now*, so age and age cohort effects collapse into one single relevant phenomenon.

Chapter 8

Experiment IV – A laugh

Laughter is a very diverse phenomenon and an integral part of everyday communication. It can be a great asset in speech synthesis because of its potential for improving naturalness. However, this depends greatly upon the appropriate choice of the shape of a given laughter instantiation. A comprehensive model predicting which laugh type should occur when would certainly be helpful (cf. e.g. Vettin, 2003; Vettin and Todt, 2004). This is, however, beyond the scope of this experiment, and there is still need for research in the basic mechanisms of laughter. To better understand some of the factors relevant for its production, and also perception, we conduct an experiment in which we suggest an articulatory scheme plausible for laughter production.[1]

We present a study on the production and perception of a laugh which originally occurred in a spontaneous dialog and therefore has an authentic context of origin. We imitate the human laugh using articulatory and *diphone* synthesis and manipulating the degree of internal variation. The laugh samples are then evaluated in listening tests. The first part of the study focuses on the acceptability (overall naturalness) of different laugh stimuli in context, whereby the stimuli vary in their degree of internal variability with regard to temporal pattern, intensity and fundamental frequency. In the second part of the study, we present the same laugh stimuli without any context, to assess their naturalness in isolation.

In the experiment on smiled vowels (Chapter 6), relevant manipulations were located in the supraglottal area of the speech tract, focusing on the creation of different target configurations of the vocal tract (smiled vowels). This was complemented by co-dependent but constant glottal features to appropriately change voice quality with larynx height. Regarding laugh synthesis, the focus now primarily lies in the glottal and subglottal area and includes the selection and control of many different gestures instead of one rather steady state of ‘smileyness’.

[1] These empirical data have been published in Lasarcyk and Trouvain (2007).

The speech production goal of this study is to develop an initial scheme to simulate a laugh with articulatory synthesis. The challenge hereby is that the exact mechanisms of human laughter production are not fully understood yet. Our artificial laughs will therefore only be approximations based on articulatory considerations and acoustic similarity.

The technical evaluation goal of this experiment is to assess the capacities of VocalTract-Lab regarding the special demands of a laugh, i.e. how a non-speech phenomenon can be addressed with and by a speech synthesizer. The special demands are characterized by extraordinarily strong variations in parameters such as lung pressure and vocal fold activity. The evaluation also addresses the performance of articulatory synthesis vs. diphone synthesis.

In terms of the overview of all experiments presented in Section 4.3, we focus on the subglottal and glottal areas since lung-pressure variations and vocal fold activity are the main parameters for creating this laugh.

In Section 8.1, we present background information on the production of laughs in humans. These considerations are used to guide the transfer into the synthesis parameters. In Section 8.2, a sample of a spontaneous human laugh is described. Its imitation with VTL is then explained, as well as the procedure of how it is being evaluated. Section 8.3 presents the evaluation results, followed by a discussion.

8.1 Production of laughs

After pointing to some general considerations regarding the production and categorization of laughs, we present in more detail the typical elements of the *song-like* type of laugh (Section 8.1.1). This is followed by a short discussion of a few articulatory aspects of its production (Section 8.1.2), which help to select useful synthesis parameters and articulatory strategies with VTL (Section 8.1.3).

Although laughs can be termed "'ha ha' laughter" (Luschei et al., 2006), it has to be emphasized that laughs are not a stereotypical sequence of laughed [ha ha] sounds (Bachorowski et al., 2001; Kipper and Todt, 2003). Kipper and Todt (2003) e.g. state that acoustic features such as fundamental frequency, intensity, and tempo (duration pattern), as well as their changing nature, "seem to be crucial for the identification and evaluation" of a laugh (Kipper and Todt, 2003: 256). When evaluating resynthesized human laughs, they found that most positive ratings were achieved by stimuli that contained varying acoustic parameters (Kipper and Todt, 2003: 267), which in their case were duration patterning (rhythm) and fundamental frequency (pitch).

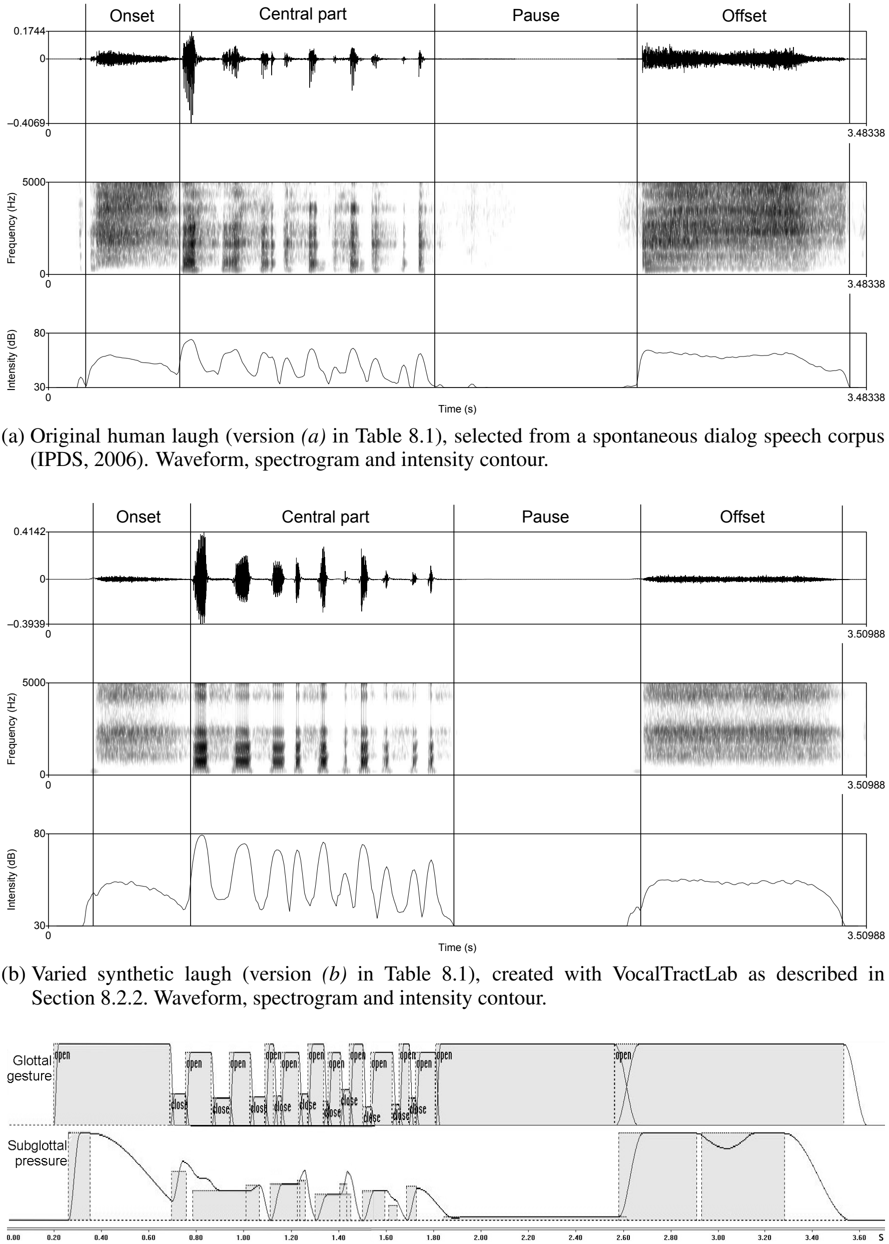
In other words, laughs are always unique, the acoustic result of the urge to laugh is never entirely predictable (Nwokah et al., 1999), and laughs can be very complex in their structure (Bachorowski et al., 2001; Chafe, 2007). This makes it very challenging to categorize different kinds of laughs and laughter in general, including "speech-synchronous forms of laughter" such as speech-laughs and smiled speech (see Trouvain, 2003, for an overview). Bachorowski et al. (2001: 1583) introduced three types of human laughs which differ on a level of segmentation that they term *bouts* (cf. Section 8.1.1). *Unvoiced grunt-like* sounds are “acoustically noisy” with “turbulence evidently arising in either the laryngeal or oral cavities.” *Snort-like* laughs are mainly unvoiced with “perceptually salient nasal-cavity turbulence.” The *song-like* type of laugh is primarily voiced and can include “comparatively stereotyped episodes of multiple vowel-like sounds with evident modulation of the fundamental frequency (f_0) as well as sounds that might best be described as giggles and chuckles.” (Bachorowski et al., 2001: 1583)

For our study, we select a song-like laugh because this laugh type has the largest internal variety of all three types. This makes it interesting to work with since it provides opportunities to test the capabilities of the synthesizer in a number of laughter production features, such as rapid glottal abduction and adduction movements (aside from regular phonation), non-linguistic frictions, vowel quality, and time-dynamic structure e.g. regarding intensity fluctuations. The other types of laughs should be studied analogously with articulatory synthesis in future work, analyzing them in spontaneous speech corpora and establishing what happens during the production process. Grunt-like laughs would presumably need friction in the pharyngeal cavity whereas snort-like laughs would additionally need friction in the nasal cavity.

8.1.1 Structure of a song-like laugh

As indicated above, laughs are very diverse in every occurrence, and can be very complex in their structure. Aside from this phenomenological diversity, the terminology to describe the features of laughs also seems rather diverse, and sometimes “confusing” (Trouvain, 2003: 2793), probably because many different disciplines have approached laughter, each from its own perspective. Coming from a phonetic point of view, we use terms related to linguistics. We are aware, though, that laughter does not behave like speech and that using linguistic terminology may make implicit claims as to the nature of some laugh features (cf. Trouvain, 2003).

Trouvain (2003) serves as the bedrock of this overview of the diverse terminology. Moving top-down through the laugh structure, the whole laugh is e.g. termed “episode” (Provine, 1993; Ruch and Ekman, 2001), “laugh response“ (Mowrer et al., 1987), “laugh event” (Luschei et al., 2006) or “laughing sound” (Rothgänger et al., 1998). It can consist of one or several “bouts” (Bachorowski et al., 2001; Ruch and Ekman, 2001) or “phrases” (Chafe, 2007), which are laugh sound sequences within one exhalation phase. The constituents of a bout are e.g. termed “laugh syllables” (Bickley and Hunnicutt, 1992), “sound bursts” (Luschei et al., 2006), “exhalations” or “pulses” (Chafe, 2007), “calls” (Bachorowski et al., 2001), “interpulse intervals” (Ruch and Ekman, 2001) or “laugh events” (Nwokah et al., 1993). From a phonetic perspective, a laugh syllable, in turn, is perceived as a sequence of a vocalic and a consonantal segment; these segments can also be considered as an alternation between sound and pause. The vocalic segment is e.g. termed “vocal peak”



(a) Original human laugh (version *(a)* in Table 8.1), selected from a spontaneous dialog speech corpus (IPDS, 2006). Waveform, spectrogram and intensity contour.

(b) Varied synthetic laugh (version *(b)* in Table 8.1), created with VocalTractLab as described in Section 8.2.2. Waveform, spectrogram and intensity contour.

(c) Excerpt of the gestural score: Glottal and subglottal tiers. The complete score is shown in Figure 8.2.

Figure 8.1: Human laugh and its articulatorily synthesized imitation.

(Nwokah et al., 1993), “vowel” (Bachorowski et al., 2001; Citardi et al., 1996; Grammer and Eibl-Eibesfeldt, 1990), “note” (Provine, 1993), “call” (Bachorowski et al., 2001), “laugh pulse” (Ruch and Ekman, 2001), “laugh burst” (Mowrer et al., 1987), “syllabic vocalization” (Mowrer, 1994), “syllable” (Provine, 1993) or “plosive” (Rothgänger et al., 1998). The consonantal segment is e.g. called “intercall interval” (Bachorowski et al., 2001) or “inter-pulse pause” (Ruch and Ekman, 2001), or described as “breathy aspiration” (Provine, 1993).

Aside from the varying terms for the agreed-upon constituents in a laugh, it is sometimes not even clear which acoustic phenomena really belong to a laugh (e.g. breathing noises at the beginning and end of a laugh). To describe the song-like laugh phenomenon studied here, we use the terms *laugh syllable*, consisting of a *laugh vowel* often alternating with a *fricative part*, i.e. the unvoiced portion of a laugh syllable or ‘laugh-/h/’ (drawn from the orthographic description of laughs as <ha-ha> or <ah-ah>). These laugh syllables constitute the *central part* of a laugh (cf. also Figure 8.1a). Some descriptions concentrate on this central part only, omitting longer silences that are typically followed by a deep and audible inhalation. Possibly, these elements provide some ongoing rhythmicity which might be important for how we perceive a laugh. Thus we extend the definition of a laugh to encompass the silences and breathing sections that frame the central part of a laugh. This also seems reasonable with regard to the capabilities of the synthesizer which can generate breathing noises as well as speech sounds (cf. Section 3.1). A laugh can thus consist of an audible forced exhalation (Luschei et al., 2006), or *onset*, followed by a *central part* with laugh syllables, a *pause*, and an *offset*, consisting of one or more audible inhalations (Chafe, 2007). These parts are illustrated in Figure 8.1a.

8.1.2 Subglottal pressure and ‘laugh vowel’

We now discuss two physiological and articulatory aspects of laughs that are helpful in developing an imitation strategy to create the synthetic laughs.

One central factor in producing laughs is the influence and the role of the subglottal pressure. The pressure level can be very different from that of speech. During speaking, we utter the phones during a prolonged exhalation phase (speech breathing, cf. Section 3.1.4), and the subglottal pressure is kept relatively steady, usually not going below a resting expiratory level (Schaeffer et al., 2002). In contrast to this, laughter often involves strongly varying pressure levels and can include extreme levels at both the low and the high end of the pressure scale: “Laughter generally takes place when the lung volume is low” (Luschei et al., 2006: 442). During a laugh the volume can go distinctly below our normal exhalation range and even below the level of functional residual lung volume, contracting our lungs to a very high degree. On the other extreme, laughter can also reach peak pressure levels of around 1.8 to 3.0 kPa (Luschei et al., 2006: 446), compared to 0.4 to 0.8 kPa in normal speech (Schaeffer et al., 2002). These extreme pressure variations are presumably an important factor when imitating a laugh event.

Another articulatory aspect of song-like laughs is the occurrence of a *laugh vowel* in the voiced parts of a laugh. It "does not appear to correspond to a standard (...) vowel" of the laugher's mother tongue (Bickley and Hunnicutt, 1992: 929). The formant patterns of laugh vowels do, however, fall into the normal range of a speaker's formant values. Complementary to that, Bachorowski et al. (2001) found that their laugh recordings generally contained "central, unarticulated sounds" (p. 1594).

8.1.3 Transfer into synthesis features

We develop a basic gestural approach to create synthetic laughs based on the characteristics of laughs described above. Each phase of the laugh is considered separately, as illustrated in more detail below. Overall, the most important synthesis parameters include the precise control of subglottal pressure, coordinated with large and fast vocal fold abduction and adduction movements. It is expected that the corresponding gestures are going to vary strongly in their amplitudes, and that a faster succession of different gestures is needed compared to 'regular' speech.

8.2 Data and analysis

A basic consideration for the generation and evaluation of a synthetic laugh is the observation that laughter is highly context-dependent. To put it with Campbell (2007b), "whom we laugh with affects how we laugh." However, not only does the communicative partner influence how we laugh with him or her, but since we also laugh following our own speech, this preceding stretch of speech also influences the laugh characteristics. For instance, Trouvain and Schröder (2004) investigated the aspect of intensity, and found that it is important to match the degree of intensity with the preceding phonetic context, otherwise the laugh would sound inappropriate. We should assume that many parameters of laughter change when a laugh occurs in different contexts. Therefore, we are going to mimic a laugh that indeed has an original context. This seems especially relevant when assessing the overall appropriateness (naturalness) of the laugh.

First, the representative human laugh sample is described (Section 8.2.1), then the basic imitation process is explained (Section 8.2.2), followed by a description of the evaluation procedures (Section 8.2.3).

8.2.1 Human laugh data

The representative human laugh of the song-like type is selected from a database of German spontaneous speech, more specifically from the dialog speech between two well acquainted adult males (IPDS, 2006). Their spontaneous discussion was elicited by a highly interactive situation where the communicative partners had to discuss the plot of an episode of a widely known German TV show (Lindenstraße). The dialog not only features speech but also a range of spontaneous nonverbal utterances such as laughter.

The structure of the selected human laugh can be seen in Figure 8.1a. It starts with an *onset*, the audible exhalation. This is followed by a *central part* with several laugh syllables in which the laugh vowel is roughly of [ɛ] quality. After that, a *pause* follows, succeeded by an *offset* with audible inhalation. The overall duration of the laugh is about 3.2 seconds. Maximum intensity varies noticeably, between about 60 dB in the onset, 65 dB in the offset, 75 dB/55 dB in the central part (first/last laugh syllables), and a silence interval of about 750 ms. The central part consists of six or seven laugh syllables (the segmentation becomes less clear towards the end).

8.2.2 Basic synthetic laugh

The representative human laugh is simulated with VTL in the following manner. First, it has to be decided which kind of gestures are most suitable for each laugh component. This is done based on the background information on human laugh production mechanisms, such as lung contractions. Once the basic gestural structure has been established, the precise durations of the individual gestures are adjusted to match the durations of the individual laugh segments in the acoustic profile of the human laugh. Finally, the amplitudes of the gestures are adjusted to match the intensity contour of the human sample.

In the following, the generation of the individual laugh elements is explained in more detail. The complete gestural score is presented in Figure 8.2. An excerpt of the two most important tiers is shown in Figure 8.1c, roughly aligned with the acoustic output of VTL as well as the acoustics of the original human laugh.

The breathing friction of the *onset* and the *offset* can be simulated by a combination of very high levels of lung pressure and a widely opened glottis. However, this only yields low to moderate levels of friction noise. To match the intensity of the friction noise more closely with the human laugh, the intensity level is increased by adding an ad-hoc consonantal gesture of a slightly constricted pharynx, based on introspection and on the observation that laughter is often produced with a tense body (cf. Section 8.1.2). This does not necessarily represent human behavior but seems a reasonable approximation.

The *central part* of the laugh is very diverse and variable but has also one major nonchanging, constant feature, namely the laugh vowel. It is kept constant across all laugh syllables to approximate the relatively small amount of supraglottal activity during the laugh. The laugh vowel has an approximate [ɛ] quality. The diversity of the central part is introduced by highly variable glottal and subglottal gestures, which induce the impression of laugh syllables. An interplay of glottal adduction and abduction gestures proves to be most convincing. The voiced portions of the laugh syllables are based on a relatively adducted glottal setting while the unvoiced portions are created using widely abducted vocal folds. The magnitude of the gestures has to be fine-tuned extensively to obtain acceptable results.

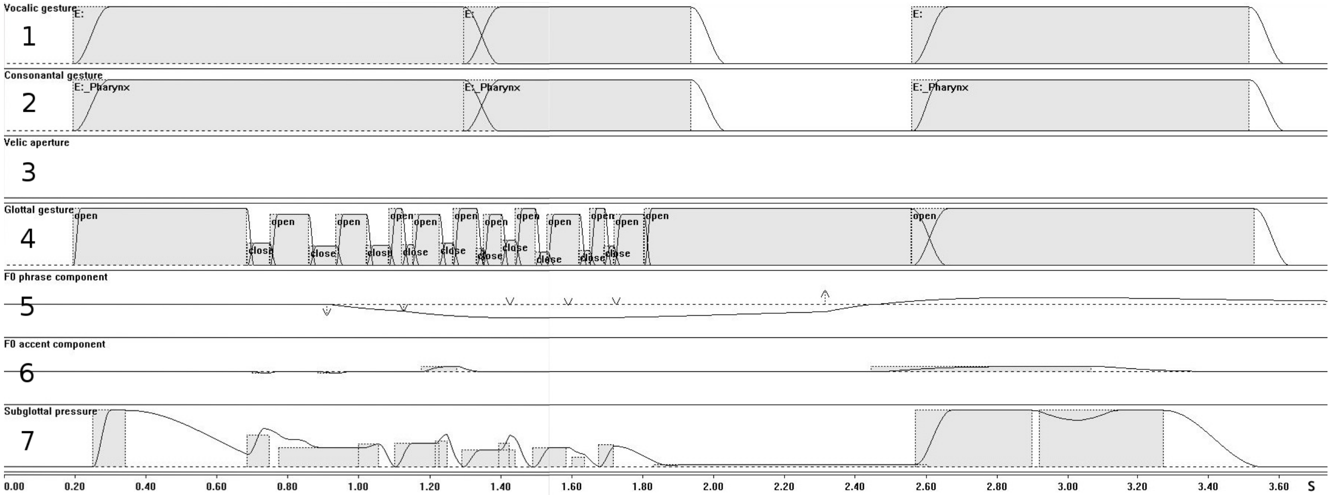
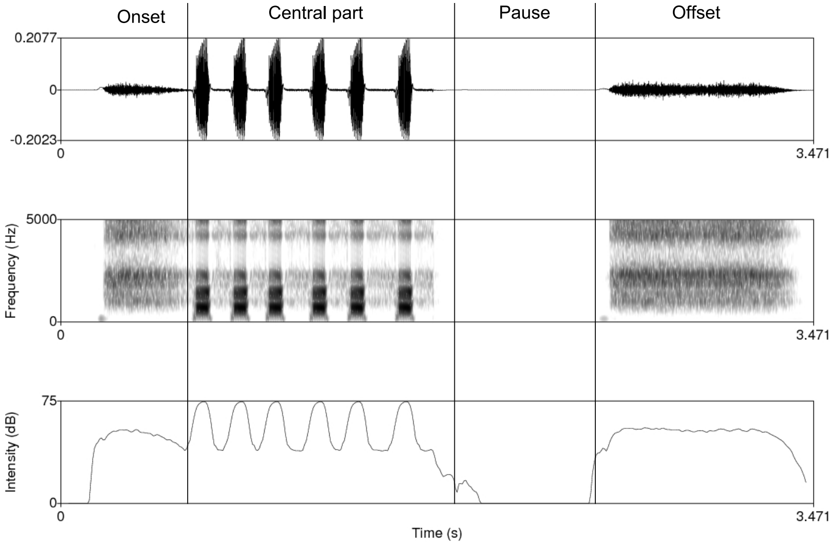
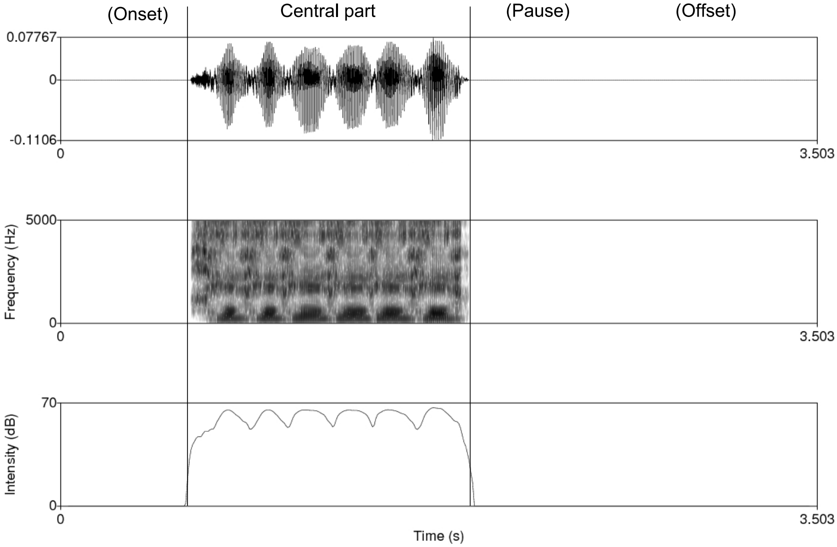


Figure 8.2: Gestural score used to simulate the detailed imitation of the laugh. It depicts the articulatory events of that laugh using the following gestures: Tier 1: Vocalic gesture “E:” for the open-mid front unrounded vowel [ɛ] underlying the whole laugh. Tier 2: Consonantal gesture “E:_Pharynx” for a slight constriction in the pharynx across the whole laugh. Tier 3: No gestures for velic aperture. Tier 4: Glottal gestures “open” and “close” in an alternating sequence, building a basis for the perceived laugh syllables. Tier 5: f_0 phrase components for the phrasal (long-term) intonation contour. Tier 6: f_0 accent components for short-term f_0 variations. Tier 7: Subglottal pressure gestures simulating the airflow throughout the laugh.



(a) Less detailed version of the target laugh (version *(c)*), created with VocalTractLab. All laugh syllables are of equal length, based on reproduced gestures from the first laugh syllable. Waveform, spectrogram and intensity contour.



(b) Stereotyped 'haha' diphone version of the target laugh (version *(d)*), created with MARY TTS (Schröder and Trouvain, 2003) using the sound segments /h/ and /ɛ/ in the laugh syllables. Waveform, spectrogram and intensity contour.

Figure 8.3: Synthesized versions *(c)* and *(d)* of the target laugh.

For example, adducting the vocal folds too much results in very prominent artifacts in the acoustic signal. The glottal alternation scheme is combined with pulmonic pressure levels adequate to produce the intensity contour of the central part of the laugh. Further increase in friction intensity is again evoked by adding a consonantal gesture of a slightly constricted pharynx, as was used in the frictions of the laugh onset and offset. f_0 gestures are added manually both for short-term variation and for the long-term course to match the perceived f_0 contour of the original.

The laugh imitation just described provides the basic scheme for the different laugh stimuli that are created for the perception test. The individual stimuli and procedures are described in the following section. The laugh that is being imitated in different versions is called the *target laugh*.

8.2.3 Laugh stimuli and perception test design

The aim of the perception test is threefold. *(i)* Find out whether articulatorily synthesized laughs are accepted as part of a natural conversation. In other words, if a laugh is placed into its original natural context, how sensitive are listeners to different forms of synthetic laughs? *(ii)* See whether the following relation holds: The more detailed the imitation of a laugh is, the better the rating of naturalness becomes when assessed in isolation. Thus, the synthetic laughs differ in the degree of detail with which they imitate the original. The more details the simulation covers, the closer it gets to the original human laugh. The manipulated features are duration pattern, intensity and f_0 contours, and the quality of the laugh vowel. *(iii)* Find out how articulatory synthesis compares to diphone synthesis. The comparison across two different synthesis techniques is carried out to see the flexibility in their performance when dealing with a non-speech utterance that has to be constructed from the available inventory, i.e. articulatory mechanisms vs. prerecorded human speech units.

In total, three different synthetic versions of the target laugh are created (versions *b, c, d* in Table 8.1 and Figures 8.1 and 8.3), two using articulatory synthesis and one using diphone synthesis. Using articulatory synthesis, we create a detailed, varied imitation based on the procedure described above in Section 8.2.2, and a less detailed version differing from the one described above in the characteristics of the central part: The first laugh syllable is simply reduplicated until the original duration of the laugh is attained, thus providing less variation within the laugh by not altering temporal, intensity, and f_0 details in the sequence of the laugh syllables. With diphone synthesis (using MARY TTS, Schröder and Trouvain, 2003), we generate a stereotyped laugh by compiling the laugh syllables from the existing speech sound inventory alternately using /h/ and /ɛ/. Since breathing sounds cannot be generated in this synthesis system, the whole laugh only consists of this central part. Lastly, the original human laugh is extracted from the dialog, representing the original version of the target laugh.

Table 8.1: Different versions of the *target laugh* of the perception test. (*art.* = articulatory.)

Stimulus name (*version*)	Voice (male)	Breathing	Central part (laugh syllables)			
			Segments	Temporal structure	f_0 contour	Intensity contour
(a) Human	Human	Exhalation and inhalation	Variation-rich natural utterance			
(b) Detailed art. imitation	Articulatory synthesis	Exhalation each time	Art. imitation with laugh vowel and laryngeal friction	Copy of original		
(c) Less detailed art. imitation				Copy and repeat of first syllable		
(d) Diphone imitation	Diphone synthesis	No breathing noises	/h ɛ/ from diphone inventory	Rough copy of original		Same for each syllable

These four laugh versions in isolation are furthermore integrated into the original dialog. The conversation is structured as depicted in Figure 8.4, the target laugh is located at the end of the short dialog, represented by the slot of 'Speaker 2', and overlapped by the laughter of Speaker 1. The slot of Speaker 2 is filled with the four different laugh versions described above, resulting in four versions of the dialog: containing *(a)* the original human laugh, *(b)* the detailed, varied imitation, *(c)* the less detailed version with the stereotyped central part, *(d)* the stereotyped laugh from the diphone synthesis system.

In total, we have thus created four stimuli consisting of four different versions of the target laugh in isolation (isolated-laugh stimuli), and four stimuli consisting of the original dialog where the different laugh versions are pasted in at the end of it (dialog stimuli). The stimuli are presented in two consecutive listening tests, using randomized orders to minimize the effects of stimulus sequence. The instructions at the beginning of the test contain a remark saying "please follow your gut feeling", emphasizing that there are no 'correct' or 'incorrect' answers. 14 subjects participate in the test (8 female, 6 male, average age 25 years, SD = 3.82 years).

In the first test, the overall naturalness of the laugh versions is tested. Subjects hear each one of the four dialog stimuli and judge the overall naturalness of the whole dialog. In the

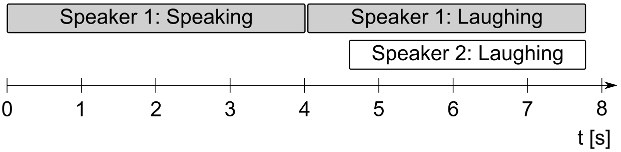


Figure 8.4: Timeline of the elements in the dialog stimuli. The contribution of Speaker 2 is the *target laugh*, represented by different versions of synthetic laughs or the original human laugh, see text.

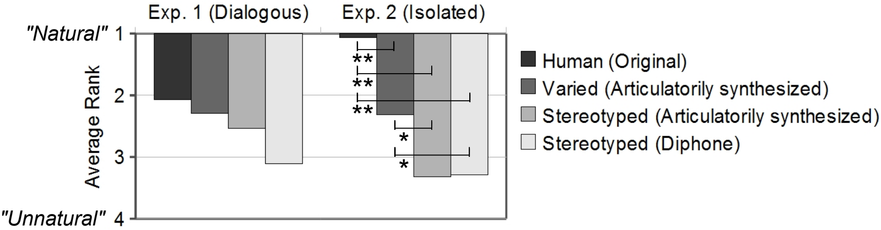


Figure 8.5: Average ranks of the dialogs containing the different laugh stimuli (left) and of the laughs in isolation (right). * indicates p $<$ 0.01, ** p $<$ 0.001.

Table 8.2: Standard deviations of the average ranks of each stimulus over all participants.

Stimulus	Exp. 1 (Dialog)	Exp. 2 (Isolated)
Human (Original)	1.00	0.65
Varied (Articulatorily synthesized)	0.66	0.73
Stereotyped (Articulatorily synthesized)	0.91	0.36
Stereotyped (Diphone)	1.19	0.36

second test, the different degrees of detailedness are tested. The subjects (the same as before) hear each one of the four isolated-laugh stimuli and are asked to rate their naturalness.

For both tests, we use the following rating scale: 1 “natural”, 2 “less natural”, 3 “rather unnatural”, and 4 “unnatural.” We use a forced-choice scale, i.e. with an even number of points, to prevent the subjects from giving ‘neutral’, indifferent judgments.

The results are processed for each experiment separately, determining the average rank of each stimulus in each of the two conditions (in dialog, in isolation). To check for significant effects of laugh type within a condition, a non-parametric Friedman test is applied (with $\alpha = 5\%$). For Test 1, we evaluate whether or not the rating of the dialog depends on the laugh placed into it. For Test 2, we evaluate whether or not the rating of the isolated laugh depends on its degree of internal variation.

8.3 Results

The average ranks of the listening tests 1 and 2 are shown in Figure 8.5. Standard deviations are shown in Table 8.2. Among the dialog stimuli, the dialog with the human laugh ranks highest (version (a) in Table 8.1), followed by the detailed articulatory imitation (b), the less detailed one (c), and the diphone version (d). However, these rankings do not differ significantly.

The ratings of the isolated laughs place the human laugh (a) as clearly more natural than all synthetic ones (mean rating of 1.07), followed by the detailed imitation (b) (2.32), and both (c) and (d) rated as rather unnatural (3.32, and 3.29, respectively). Overall, these ratings differ significantly, and post-hoc pairwise comparison (Wilcoxon, with Bonferroni correction to a significance threshold of 0.83%) shows a significant difference between all pairs but (c) and (d). In other words, the human version is ranked as significantly more natural than all synthetic ones ($p < 0.001$); and the detailed imitation (b) is ranked as significantly more natural than the less detailed and stereotyped versions (c) ($p = 0.002$) and (d) ($p = 0.008$), regardless of the synthesis system used.

8.4 Discussion

In this experiment, we suggested articulatory patterns plausible for laughter production which were implemented in articulatory speech synthesis. The degree of internal variation of a target laugh was manipulated, and in a perception test the laughs were compared to the human original and a diphone synthesized version.

8.4.1 Phonetic aspects

All versions of the target laugh were rated as equally natural when heard in the context of the original human conversation, i.e. no significant differences in the ranking were found. In particular, the artificial laughs in the dialogs did not significantly decrease the naturalness ratings compared to the dialog containing the natural laugh. This can be interpreted as providing a positive answer to the question whether articulatorily synthesized laughs would be accepted in a natural conversational setting. The initial scheme to imitate the different elements of a laugh therefore seems usable.

Regarding the second question in this experiment, related to the internal variation of the stimuli, we found that the more detailed the imitation was, the better the scores were for perceived naturalness. Since human speech varies strongly as well, this seems to be a plausible answer, and a higher degree of variation, of course, comes closer to what humans produce and thus goes hand in hand with naturalness. The finding confirms results in Kipper and Todt (2003), where a variation in the acoustic parameters rhythm and pitch yielded the most positive ratings. In our experiment the two laughs synthesized with VTL differed with respect to variation in duration, intensity and f_0 in the sequence of the laugh syllables.

We evaluated the same target laugh both in isolation and context-based (in a dialog). While the isolated laughs received significantly different ratings, the in-context laughs seemed to function more homogeneously. This indicates that if a well-matching laugh is presented in its original context, it is more likely to be accepted in a dialog, even if it is synthetic, than when it is presented in isolation, where the unnaturalness of synthetic laughs is more prominent. Evaluating a laugh both in isolation and in context proved to be a good means of assessing both the articulatory details of the imitation as well as its overall appropriateness. Thus, both evaluation setups have particular advantages and can complement each other. This is briefly discussed here.

Since real instances of laughter always occur in some sort of context, it can be argued that evaluating laughter in isolation is a rather unnatural approach because it resembles a speech-lab setting. However, this setup can nevertheless provide insights into the perception of laughter. An isolated laugh requires a certain amount of abstraction on the part of the listeners and at the same time enables them to pay attention to details otherwise masked. This way their ratings can indicate e.g. which individual *details* contribute to a laugh being more natural.

Presenting laughter in context reveals how flexible listeners are in what they accept in terms of micro-features of a laugh as long as the macro-features determining the appropriateness for that context are acceptable. The overall naturalness ratings are higher, presumably because they reflect the fact that everything fits together well rather than being an assessment of the articulatory details. In context-based evaluations, the manipulation variables cannot be separated as easily as in isolation-based evaluations. Some variables may also be masked by overlapping vocalizations from the conversational partner – but this is also true in the real world. Thus, in principle, the evaluation of laughter in context has the advantage that it can lead to more discourse-related knowledge to design *overall* appropriate synthetic laughs (see e.g. Trouvain and Schröder, 2004).

8.4.2 Synthesis technical aspects

The experiment carried out in this chapter represents a first exploration of VTL into laugh synthesis, and also compares its performance with diphone synthesis. While the basic articulatory ideas could be adequately executed in VTL, some limitations in glottal and subglottal control were found, making it evident that the synthesizer was originally designed for ‘regular’ speech. This is discussed here, followed by considerations regarding perceived naturalness, and a short account on the required sensitivity regarding the articulatory control of laughs. Finally, aspects across different synthesis techniques are discussed.

Regarding limitations in glottal control, the increased demand on glottal activity was hard to manage with the synthesizer, and the newer version (see Chapter 3, p. 41) seems even slightly less flexible in controlling the glottal mechanisms at such a high speed and with such high amplitudes. Regarding limitations in subglottal control, the peak lung pressure was effectively too low for laugh synthesis. This was compensated by the ad-hoc pharyngeal constriction gesture which created an additional, supraglottal noise source. Furthermore, breathing could only be simulated by exhaling, not by inhaling, as it often occurs towards the end of laughs. However, we were able to use *ex*haling in both occurrences because of its acoustic similarity to inhaling. These breathing-related issues need further basic research with humans first. Then, a slightly more flexible implementation of breathing in VTL would be helpful for future laugh synthesis.

The audio rendering did not yet sound very natural. This is presumably due to the design of the gestural scores and the current acoustic simulation of the synthesizer. Firstly, the gestural scores presented here can only be viewed as the current best fitting solutions to imitating laughter. They may represent one possible way of simulating what happens when a human produces a laugh but they do not claim to completely reflect human production processes. Therefore, the acoustic quality might sound a little strange. Secondly, the acoustic simulation in the synthesizer creates somewhat artificial sounding utterances, regardless of the type of utterance. With 'regular' speech, a listener can use their own expectations depending on the context, thereby increasing ratings of intelligibility, and naturalness. With a laugh however, we do not have any typical linguistic expectations, so the details of the utterance may be more important than those of 'regular' words. Therefore, achieving really natural sounding laughs seems a very hard task. Moreover, even human laughs, evaluated in isolation, may sound unnatural because usually we hear laughs within a certain context.

The following two considerations are pointed out to show that one main challenge in the laugh imitation task lies in the combination of sensitive articulatory controls with the largely unexplored use of articulatory mechanisms with respect to simulating laughs. Therefore, the proposed articulatory solutions for the laugh samples are a result of informed trial-and-error iterations, and they only represent one local solution to the problem of laugh simulation. Firstly, the varied pattern of alternating adduction and abduction gestures in version (b) is the result of an extensive fine-tuning process dealing with different magnitudes of the gestures. Small changes in the gestural score resulted in gross changes in the acoustic output. This illustrates the sensitivity of the control of the synthesis mechanisms. Secondly, the whole output improved strongly when the pulmonic pressure was decreased considerably, exemplifying how a simple change can result in a totally different output quality.

Across synthesis techniques, each of the two techniques used here has shown some advantages. Although the diphone imitation (d) was merely based on a concatenation of 'regular' speech sounds, it apparently matched the dialog environment (Test 1) just as well as the rough articulatory imitation (c). The diphone laugh might have had the advantage of stemming from a human voice in the first place, i.e. during database recording, preserving essential traits of naturalness. The strategy of concatenating /h/ and /ɛ/ worked well for song-like laughs, which have been described as being "in some ways (...) speech-like" (Bickley and Hunnicutt, 1992). However, flexible synthesis of grunts or snorts hardly lies within the scope of a concatenative synthesis system that was designed for speech sounds alone. Articulatory synthesis, on the other hand, proved to be more natural in the isolated-laugh condition (Test 2) presumably because it can imitate the laugh in more detail by drawing on the different articulatory mechanisms provided by VTL, and it is possibly also capable of imitating grunts and snorts.

This experiment shows that a laugh, and in principle other affective expressions, can be synthesized with different synthesis techniques, offering different advantages: Inherent naturalness of the voice vs. flexibility in the structure of the laugh. To sound appropriate, the utterances have to be adapted to the context. With diphone synthesis based on a speech database of read speech, the variation possibilities are limited and depend on suitable corpus items. When prerecorded laughs are used, they might sound misplaced or inappropriate when actually synthesized in a specific context (Trouvain and Schröder, 2004). Post-hoc acoustic manipulations are possible to some extent but the character of the laugh is predefined in the corpus. Nevertheless, one can still explore such things as the effects that laughter has on the listener by directly manipulating stimuli on the acoustic, or signal level (see e.g. Sundaram and Narayanan, 2007; Campbell, 2006; Trouvain and Schröder, 2004). However, one cannot automatically synthesize such a wide range of laughs as when articulatorily imitating the speech production processes of laughter.

Chapter 9

Experiment V – A speech-laughed word

The present chapter deals with the imitation of a speech-laugh. A speech-laugh, or speech-laughing, is characterized by the simultaneous production of laughter and, usually, still intelligible segmental content. It seems to occur very often in spontaneous speech but may often be hard to capture due to its complex, and sometimes subtle, acoustic nature. Depending on the circumstances, one or the other 'modality' – i.e. speaking or laughing – may be more dominant or may be the primary way of intended communication: On the one hand, speech can be the primary means of communication in situations where we talk to someone, and a sudden urge to laugh, or laughter as a non-verbal complement is incorporated into the speech. On the other hand, we may encounter situations in which our primary means of communication is a laugh, and we want to add words to it. In this way, both 'modalities' may influence each other to different degrees until the typical balance between speaking and laughing is reached.

As with smiled speech, speech-laughing represents a means of displaying affective information and a linguistic message at the same time. The exact mechanisms of speech-laughs are not fully understood yet. So to better understand some aspects of its nature, we investigate salient features of a speech-laugh with articulatory synthesis. Furthering this understanding could also contribute to making a synthetic voice appear more natural.

In the following experiment, we suggest an articulatory scheme for imitating a one-word human speech-laugh with articulatory synthesis by presenting plausible articulatory patterns. We create different speech-laugh imitations, which vary with respect to the use of syllabic pulsation and a smiled laugh-vowel. Perceptual evaluation is carried out to determine the effects of these two manipulation variables on perceived amusement and naturalness. This experiment can be regarded as a follow-up on the laugh imitation presented in Chapter 8. Additionally, experiences gained from the imitation of smiled vowels (Chapter 6) are integrated.

The speech production goal of this experiment is to gain some insight into the mechanisms relevant for the combination of speaking and laughing by developing this initial scheme of how to simulate the articulatory details of a speech-laugh.

The technical evaluation goal is similar to the one in the previous experiment, where an assessment was made of how well VTL can deal with the special glottal and subglottal demands during laugh synthesis. This is now complemented by evaluating how additional supraglottal demands, arising from the linguistic content of the utterance, can be integrated.

With respect to the overview of all experiments (Section 4.3), we focus here on the subglottal, glottal, and supraglottal area of speech production.

In Section 9.1, background information is presented on the production of speech-laughs, from which relevant synthesis parameters are identified. In Section 9.2, we describe a human speech-laugh sample, its synthetic imitation, and the speech-laugh evaluation procedure. Results are presented in Section 9.3, followed by a discussion.

9.1 Production of speech-laughs

Some basic considerations about the nature of speech-laughs are discussed here. The diversity of laughed speech makes it very challenging to explore. On the one hand, we are faced with the variations of spontaneous speech which are characterized by highly variable supraglottal actions. On the other hand, we are faced with the great variety that is found in laughs. Here, the emphasis in articulation is on glottal and subglottal (i.e. pulmonic) mechanisms, with fewer demands being placed on supraglottal activity than in 'regular' speech (see Table 9.1). These different foci of activity in speaking vs. laughing seem to make it possible to merge the one with the other, resulting in speech-laughs (Chafe, 2007).

Table 9.1: Focus of activity in speaking, laughing, and speech-laughing.

Area	Speaking	Laughing	Speech-Laughing
Supraglottal	**Varying**	Less varied	**Varying**
Glottal	Less varied	**Varying**	**Varying**
Subglottal	Less varied	**Varying**	**Varying**

However, to speech-laugh does not mean to simply superimpose laughter onto speech (Trouvain, 2001), instead it seems to involve a deep reorganization of both systems. Nwokah et al. (1999) put forward the possible mutual influences from speech and laughing by proposing a model within a 'dynamic systems' perspective. Every outcome of the merging of the two systems is unique and not entirely predictable. In most cases, neither one of the systems is dominant over the other, it is rather a "delicate balance" between the two. This is not the case, though, when a speaker is e.g. faced with a really forceful laugh; then the dominance of the laugh system is substantially higher than the one of the vocal, i.e. speech system, and the vocal apparatus is apparently being taken over by the laugh system.

Before going into more details of the structure of speech-laughs (Section 9.1.2), we outline the occurrence frequencies of speech-laughs found in different studies (Section 9.1.1). Finally, relevant synthesis features are deduced (Section 9.1.3).

9.1.1 Proliferation of speech-laughs

Nwokah et al. (1999) state that the occurrence of laughed speech depends greatly on the context and the individual. The authors recorded speech in mother-child free play situations and found it to contain a large amount of speech-laughs. Infant-directed speech is said to be an "ideal" setting to gather speech-laughs. The question arose whether this could be transferred to different interaction partners, especially ones of adult age, who are not so intimately related to each other as a mother to her child.

As a sample of adult-adult communication, we searched for speech-laughs in two corpora of German spontaneous dialogs (IPDS, 1995-1997, 2006). Both corpora consist of speech recordings made in highly interactive situations. One deals with appointment making, the other one is the same as introduced in Section 8.2.1, with discussions about an episode of a widely known German TV show. Although it is adult-adult communication in all dialogs, the levels of intimacy vary substantially. The German TV show is discussed by pairs of friends, whereas in the appointment making scenario the dialog partners are only superficially acquainted. However, the level of intimacy is never as high as in the mother-child corpus mentioned above.

Although the recording situations are identical across all dialogs within each corpus, we find substantial differences in the number of speech-laughs that occur in each pair's dialog. In the conversations, which are of a length of about 10 to 15 minutes each, the overall number of speech-laughs ranges from 2 or 3 in one dialog to over 20 or 30 instances in another. This supports the claim that the number of occurrences of speech-laughs depends greatly on the social context and the individual (see Nwokah et al., 1999, and e.g. Campbell, 2007a,b).

The above observations contrast with the findings of Provine (1993), who studied the dialog speech of college students and found a close to zero percentage of speech-laughs ("Interruption" kind), compared to the amount of laughs found. For documentation, a system of live coding was used to count the laughter instances. In other words, the perceived events were marked on-line, without recording the speech itself. We would assume, from our own listening experience, that the way of collecting those observations might not have been perfectly suited to a phenomenon so elusive as a stretch of laughed speech. Even when inspecting *recorded* speech, which can be scrutinized in depth, we find it very challenging to determine whether some globally 'unusual' stretch of speech is a laugh, a speech-laugh, smiled speech, or only a strong breath.

The evidence from spontaneous speech corpora indicates that laughed speech occurs frequently, at least in certain settings. In addition, it seems to be quite an elusive phenomenon which can easily be overlooked.

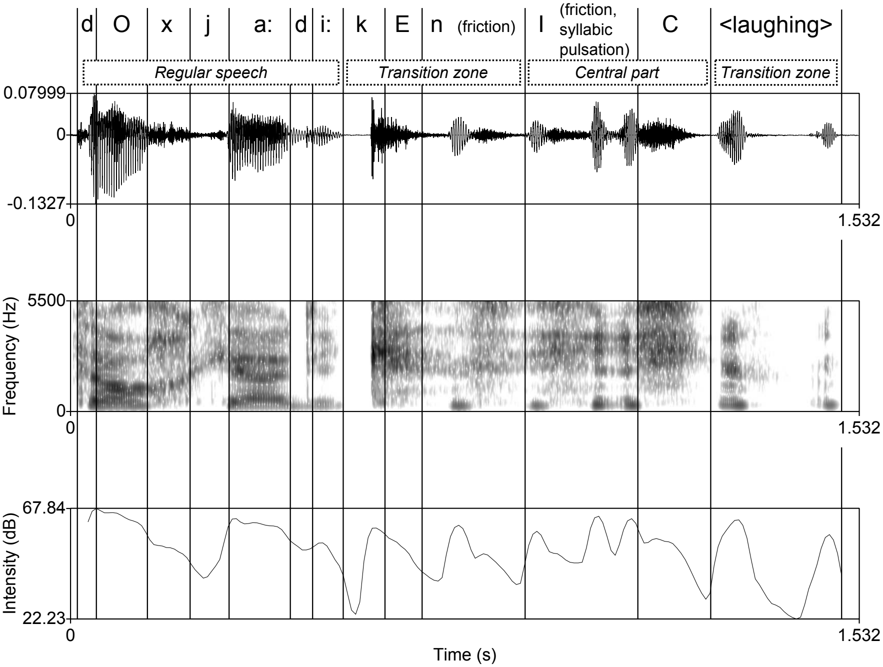


Figure 9.1: Human speech-laugh, with frictions and syllabic pulsation, ending in a minor laugh. *<Doch, ja, die kenn' ich.> (<Yes, true, I do know her.>)*

9.1.2 Features of speech-laughs

In this section, we describe the elements that can be found in speech-laughs. Firstly, an illustration of a basic typical speech-laugh structure is given. Then, the acoustic characteristics are described in more detail, presenting an account by Nwokah et al. (1999), which discusses the possible origin of the different speech-laugh elements. This is complemented by characteristics mentioned in Trouvain (2001) and Chafe (2007) regarding vocal tremor, and the mention of nasality in Nwokah et al. (1999). Finally, these accounts are complemented by a description of the speech-laugh phenomena occurring in the speech corpora IPDS (1995-1997) and IPDS (2006).

Basic structure

Referring to the terminology introduced for laughs in Chapter 8, a speech-laugh also has a *central part* whose features are described below. It can be surrounded by speech or other emotive phenomena, connected to the speech-laugh by short transition zones, leading into the speech-laugh and trailing out of it, e.g. with raised f_0 or emerging friction noises (cf.

Figure 9.1). These features could also be counted as central parts of the speech-laugh but it is not trivial to precisely determine the boundaries. It seems reasonable to assume a period of *transition* between different emotional states, since we are not discrete beings jumping instantaneously from one emotional state to the other.

Effects from speech, laughing, and individual variability

As pointed out at the beginning of this chapter and in Section 9.1, the mechanisms of speaking and laughing both seem to influence the final shape of a speech-laugh. We now describe two complementary perspectives to characterize speech-laughs in more detail, firstly by regarding laughing as the primary component, and how speech influences it, secondly, by regarding speaking as the primary intention, and which laugh features influence it. Afterwards, we discuss aspects of variability.

Following the observations found in Nwokah et al. (1999), the most important influence from the *speech side* is that vowels and consonants are produced. Thus, the previously introduced laughter-induced 'laugh vowel' (see Section 8.1.2, p. 115) is replaced by the different speech sounds needed for the linguistic message. Also, the fundamental frequency is largely, but not always, adopted from speech. Lastly, speech seems to increase the duration of laughter, presumably as a way of allowing both speech that is still intelligible and the rhythm of laughter.

The most salient influence from the *laughter side* is probably the characteristic repetitive rhythm. It manipulates the speech by leading to vowel elongations and, most of all, syllabic pulsation. The latter seems to be one of the most recognizable features of speech-laughs. Furthermore, the typical unvoiced sounds of a laugh syllable are inserted (a glottal stop or friction-[h]). Most of the times, the amplitude contour is taken from the laughter side, and sometimes breathiness is introduced into the speech sounds due to the laughing activity. These modifications are highly variable, but they lead to the same overall percept of a speech-laugh.

While identifiable, stable components of both laughter and speech are retained (stability factors), we find extensive variability in individual speech-laugh styles (Nwokah et al., 1999: 891f). The set of 'stable characteristics' comprises e.g. physiological constraints in expiratory length as well as what is termed "spikiness" (p. 892): a quick and extensive amplitude modulation directly related to syllabic pulsation. Variability is generated by different strategies of balancing respiratory and laryngeal demands, either by expanding words by syllabic pulsation, or by compressing them. In three-word or longer speech-laughs, elisions as in very fast speech are found (Nwokah et al., 1999). Further variability factors include the timing of the onset of laughter within the speech fragment, and whether or not breathiness occurs. These variability factors represent skills that evolved during communication, and they constantly adapt to the current context. In sum, they make up the idiosyncratic pattern of an individual's speech-laugh.

To a large extent, the above account matches the descriptions put forward by Trouvain (2001), where the analyzed speech-laugh tokens mostly show a “reinforced expiratory activity” (p. 636). This results e.g. in breathy vowels, inserted or amplified aspirations in plosives and fricatives, and devoiced nasals. These kinds of occurrences could be described as *fricativized* speech-laughs. Furthermore, pitch is increased and speech-laughs mostly span two syllables (Trouvain, 2001: 635). Nwokah et al. (1999), however, claim a common duration of about two words. Given the great variability, the findings are probably compatible. Another feature of variability is the finding that speech-laughs can occur in any position within a phrase (Trouvain, 2001), with no predominance of function or content words (Nwokah et al., 1999). Both studies claim that most speech-laughs start or end simultaneously with speech articulation.

Vocal tremor

In some kinds of speech-laughs we find vocal tremor or vibrato during voiced segments, especially vowels. While Trouvain (2001) states that this phenomenon occurs only sometimes in laughed speech, Chafe (2007: Chapter 3) e.g. describes it as occurring often. Vocal vibrato, which “may be a cultivated vocal tremor” (Titze, 1995: 704), is used voluntarily in singing to enhance the quality of the sound (Titze et al., 2002). Vocal tremor, on the other hand, is associated with involuntary frequency modulations, is more irregular and is considered pathological (Titze, 1994; Kreiman et al., 2003).

In the environment of an emotional utterance such as a speech-laugh, the frequency modulation of the fundamental frequency appears to be ‘uncontrolled’. Thus, the term ‘tremor’ might be more appropriate here.[1] The origins of the rhythmic oscillations of the vibrato and tremor, how they are triggered and sustained, are not fully understood yet (see e.g. Leydon et al., 2003, for an introductory overview), in particular with regard to (tremulous) speech-laughs.

To characterize the strength of vocal tremor in speech-laughs, we can use a comparison to the irregularities commonly found in voiced speech: Acoustic analysis will detect slight modulations in *any* voice, and changes in fundamental frequency at a rate of between 2 and 12 Hz are not perceived as unusual. However, vocal tremor or vocal vibrato seem to be more pronounced than that, and thus are perceived as a prominent feature in the voice.

Nasality

The last phonetic aspect of speech-laughs to be mentioned here is the feature of nasality. Nwokah et al. (1999) exclude it due to low inter-annotator agreement. We note a similar perceptual difficulty and therefore exclude this feature in this experiment. Nevertheless, we

[1] Despite the uncontrolled tremor, the general articulation is still controlled on a subconscious level, adjusted to the rules of cultural display (Ekman, 1977) and to the context of the situation. However, due to the competing demands on the vocal organs, the control is made more difficult.

recognize an issue with nasalized sounds, since auditory analysis clearly hints at unusual nasal activity during speech-laughs, only it is hard to capture. Future work could investigate reliable measurements to find out what role nasality plays in laughed speech.

Our own corpus investigations

Complementing the descriptions above, our own analysis reveals the following speech-laugh features in the two German corpora of spontaneous speech (IPDS, 1995-1997, 2006). In a majority of instances, we find added and/or enforced frictions. We count a considerable number of perceived tremulous speech-laughs, but some of them appear not to be a true fundamental frequency tremor on an individual vowel. They rather consist of a pulsating sequence of consonants and vowels, merging into what we regard as syllabic pulsations. The results are 'reduplicated' syllables as in laugh syllables (<yea-hea>) or very fast articulation rates adapted to a pulsating intensity contour (see also Nwokah et al., 1999). We also note simultaneous smiling in the voice quite often, as well as some impressions of nasality. The laughed stretches of speech often, but not always, show increased intensity levels. Regarding the context, the laughed speech is predominantly combined with laughs, audible inhalations, and stretches of smiled speech.

Although it is difficult to identify a representative pattern of speech-laughs, a – tentative – subcategorization into different kinds of speech-laughs in our corpora is proposed here. A rough distinction can be made between *fricativized* and *tremulous* speech-laughs, although the tremulous ones seem to be less frequent in our data. We find indications that the number of occurrences might depend largely on which style an individual speaker prefers. As it was not the target of our investigation to fully analyze, distill and categorize the occurrences of laughed speech in these speech corpora, detailed accounts are left for future work. The initial inspection is carried out to provide a rough idea about what could be representative samples for articulatory resynthesis, and the distinction between fricativized and tremulous speech-laughs serves as an initial guide to find resynthesis candidates.

9.1.3 Transfer into synthesis features

A comparison of the articulatory characteristics of speech, laughing and speech-laughing helps to find the necessary synthesis features.

While laughs put a strong focus on the subglottal and glottal areas (see Chapter 8), speech-laughs cover virtually every area of the vocal apparatus, as indicated above. The basic difference between synthesized 'regular' speech, laughing, and laughed speech can be described as follows (cf. Table 9.1, p. 128). When synthesizing 'regular' speech, we use a sequence of varying vocal tract shapes to produce the consonantal and vocalic sounds. The emphasis thus lies on supraglottal variation complemented by a specific set of glottal gestures when e.g. plosives are involved. Other gestures, such as f_0 control, voice quality, or lung pressure, can be kept relatively stable. For laughs, the focus of activity and variation is located on glottal and pulmonic gestures (strong vocal fold abductions/adductions, f_0 control, lung pressure), while the supraglottal configuration is more or less constant with an only gradually changing 'laugh vowel'. Finally, to create speech-laughs, we need a high degree of activity in all the above areas: Not only do the subglottal and glottal gestures, which are laugh-induced, vary quickly over time. The supraglottal and glottal gestures, which are segmentally motivated, have to be adjusted appropriately as well when generating speech-laughs. For instance, increased friction levels are needed, and prolonged or shortened sound segments, since the laugh profoundly alters the speech across which it stretches. However, despite all modifications, segmental intelligibility needs to be retained.

Based on these considerations, we first build a gestural score to represent an 'emotionally neutral' (linguistic) version of an utterance. This basic segmental content needs to stay intact for the speech-laugh to be intelligible. The initial utterance is gradually modified into a speech-laugh by implementing two of the typical features mentioned above: syllabic pulsation, which might also induce the perception of vocal tremor, and smiled vowel quality. Syllabic pulsation is implemented by glottal and subglottal gestures. Smiled vowel quality is obtained by supraglottal manipulations.

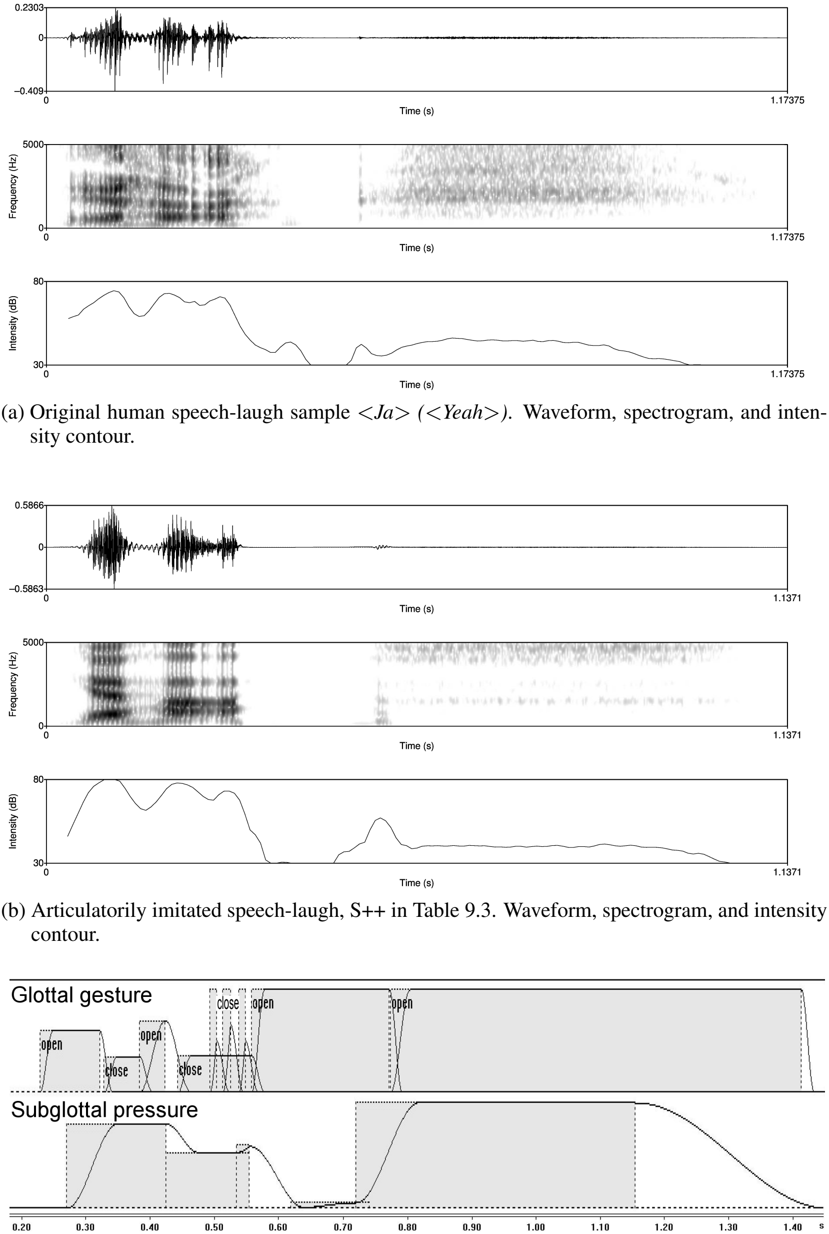
9.2 Data and analysis

First, the representative human speech-laugh sample is described (Section 9.2.1), which is then imitated with VTL (Section 9.2.2). Finally, we present the evaluation procedures (Section 9.2.3).

9.2.1 Human speech-laugh data

While searching the databases (IPDS, 1995-1997, 2006) to find a representative human speech-laugh for resynthesis, it becomes evident that only the short instances of the speech-laughs in our corpora are suited for resynthesis. The longer an utterance is, the more complicated the gestural score will be even for the 'emotionally neutral' version because it has to incorporate all the segmental (linguistic) information in the gestural alignment. This presents a problem of fine-tuning which is not the focus of this experiment but is brought up in a later experiment (Chapter 11).

Due to these restrictions on segmental complexity, we select a one-word phrase that despite its short duration (about 1100 ms) shows the two prominent speech-laugh features selected above, i.e. syllabic pulsation and smiled vowel quality. Its acoustic characteristics are depicted in Figure 9.2a. The wording of the selected utterance is [jaː] <ja> 'yes', spoken by a male adult, with the *central part* of the speech-laugh being located mainly on the vowel. Syllabic pulsation shows itself in the reduplication of the nucleus [aː], accompanied by a syllable border that sounds similar to a strongly fricated [h] (friction-h). This considerably prolongs the vowel. The quality of the [aː] is slightly raised (i.e. not fully open), inducing



(a) Original human speech-laugh sample *<Ja> (<Yeah>)*. Waveform, spectrogram, and intensity contour.

(b) Articulatorily imitated speech-laugh, S++ in Table 9.3. Waveform, spectrogram, and intensity contour.

(c) Excerpt of the gestural score: Glottal and subglottal tiers.

Figure 9.2: Human speech-laugh and its articulatorily synthesized imitation.

the auditory impression of a smiled setting of the vocal tract. The *central part* of this sample speech-laugh is followed by an inhalation phase.

9.2.2 Synthetic speech-laugh data

When creating the gestural score to simulate the selected speech-laugh, we build upon the strategies used for the laughter simulation in the previous chapter. The initial gestural score, however, is built without any laugh-like elements. It represents the linguistic or 'neutral' version of the target phrase [jaː].

To create the complete speech-laugh, smiled vowel quality and syllabic pulsation are added in the following manner: First, the vowel quality in the basic version of [jaː] is changed into a smiled quality. It is created manually in the GUI of VTL by horizontally retracting the lips of the default [aː] phone (parameter LP, cf. Table 3.1), and by slightly raising the lower jaw until the vowel has an audible 'smiled' quality (parameter JA). The formant values of the smiled and the default [aː] are presented in Table 9.2.

Table 9.2: Neutral vs. smiled laugh-vowel quality. Formant frequencies (in Hz) of the vowels used in the speech laugh.

Vowel	F_1	F_2	F_3
$[a]_{neutral}$	792	1311	2467
$[a]_{smiled}$	860	1385	2599

In a second step, the score with the smiled [jaː] is enhanced and prolonged by adding syllabic pulsation. This is done by placing several abduction gestures onto the glottal tier. Additionally, creak is inserted during the syllabic pulsation since this is also audible in the human sample. This is achieved by adding a sequence of short and intensive adduction gestures towards the end of the vowel (cf. Figure 9.3).

Finally, the f_0 contour and lung pressure behavior are adjusted in detail to match the original speech-laugh. The resulting speech-laugh is labeled *Stimulus S++* in the matrix shown in Table 9.3 because both speech-laugh features are active in this imitation. This 'fully active' speech-laugh is depicted in Figure 9.2b, aligned with the human speech-laugh, and complemented with an excerpt of the gestural score. The complete gestural score is shown in Figure 9.3. It represents the imitation which is most closely related to the original human speech-laugh sample (stimulus H++ in Table 9.3).

9.2.3 Speech-laugh stimuli and perception test design

The gestural score of the 'fully active' speech-laugh (version S++, described in Section 9.2.2), is altered in two features to create the material for the perception test: Presence or absence of syllabic pulsation and presence or absence of smiled vowel quality. This results in four synthetic stimuli (S), depicted as a matrix in Table 9.3 and complemented by the original

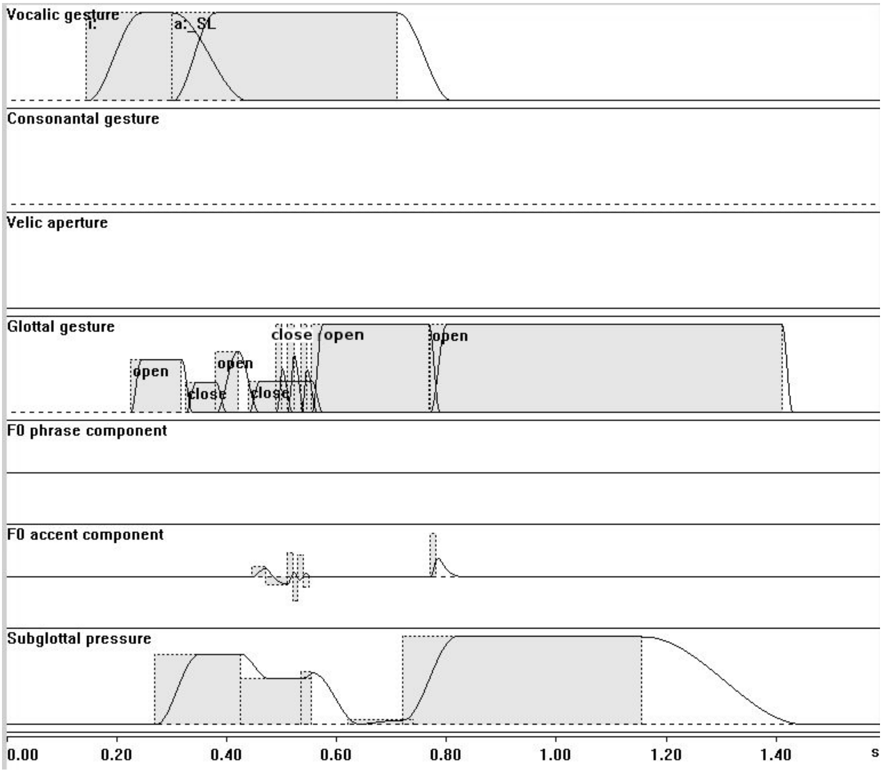
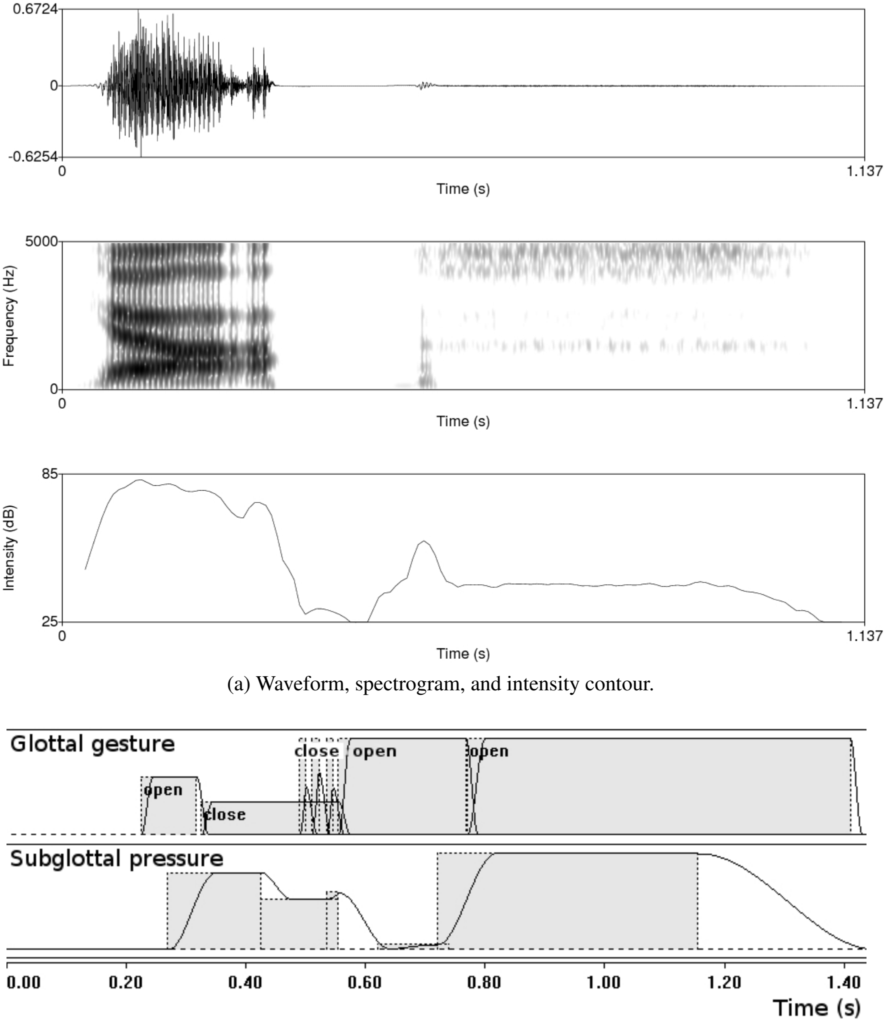


Figure 9.3: Gestural score of the 'fully active' speech laugh, S++ in Table 9.3, featuring smiled vowel quality ([a:_SL] on the vocalic tier) and syllabic pulsation (glottal and subglottal tiers). Its acoustic characteristics are shown in Figure 9.2b.

Table 9.3: Matrix of stimuli for the speech-laugh perception test. The first character indicates human (H) or synthetic (S) voice, followed by an indication of presence or absence (+/–) of syllabic pulsation and smiled laugh-vowel quality.

Stimulus	Syllabic pulsation	Smiled vowel quality
H++	Yes	Yes
S++	Yes	Yes
S+ –	Yes	No
S– +	No	Yes
S– –	No	No



(a) Waveform, spectrogram, and intensity contour.

(b) Excerpt of the gestural score: Glottal and subglottal tier. Roughly aligned with the acoustic information above.

Figure 9.4: Articulatorily imitated speech-laugh without syllabic pulsation.

human speech-laugh (H) as a fifth stimulus. As mentioned above, the acoustic characteristics of the ‘fully active’ speech laugh S++ are depicted in Figure 9.2b. The second stimulus with syllabic pulsation (S+ –) looks virtually identical and is not depicted separately. In contrast to the ‘fully active’ speech-laugh, it does not feature smiled vowel quality. The general acoustic structure and relevant gestures of the two synthetic speech-laughs *without* syllabic pulsation (S– + and S– –) are shown in Figure 9.4. Both stimuli without syllabic pulsation look virtually identical, their only difference again being the presence or absence of the smiled vowel quality.

The listening test addresses the following questions: *(i)* Does the feature distribution (see matrix in Table 9.3) influence the degree of perceived amusement? *(ii)* Do the features influence the degree of perceived naturalness? Further on: Are *(i)* and *(ii)* correlated, and where on an absolute scale are the ratings of naturalness located? The last question should clarify what ‘sounding more natural’ actually means: Where is the synthesis system located in the continuum of naturalness regarding these stimuli?

The four synthetic speech-laughs and the original recording are presented in a pairwise preference test to obtain an average ranking consisting of five ranks (first to least preferred stimulus). The stimuli are paired up in all permutations, i.e. both possible sequences of each pairing and same-same pairings. The sequence of stimuli is randomized to minimize ordering effects. After playback of a stimulus pair, we ask the participants to mark the utterance that they perceive as coming from “the more amused speaker.” Then we play back the pair of stimuli again, asking which one of the two sounds more natural, with a follow-up question as to how natural the “more natural one” sounds on an absolute scale from 1 “very natural” to 5 “really unnatural.”

The experiment is carried out in two group sessions, with a total of 25 participants (19 female, 6 male, average age 23.9 years, SD = 6.4 years). The answers are collected leaving out the ratings for same-same pairings. Two participants are excluded since they did not fully comply with the forced choice task. Thus, the individual rating profiles of 23 participants are compiled and analyzed.

9.3 Results

The accumulated responses to the question “Which features lead to a high degree of perceived amusement?” are depicted in the ranking shown on the left side of Figure 9.5. The distinctions the subjects make are overall highly significant (Friedman test: $\chi^2 = 49.83$, $p < 0.001$). Post-hoc pairwise comparison (Wilcoxon, Bonferroni correction to a significance level of 0.5 %) shows that the original stimulus (H++) is rated as sounding significantly more amused than any of the synthetic stimuli. Additionally, S++ (synthetic, with both syllabic pulsation and smiled vowel) sounds significantly more amused than S– + and S– –, neither of which feature syllabic pulsation. This seems to indicate that syllabic pulsation can increase the degree of perceived amusement. Between S++ and S+ –, both featuring

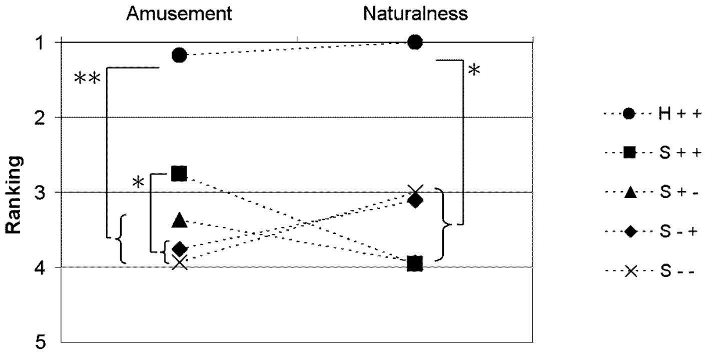


Figure 9.5: Average rankings for speech-laughs with respect to amusement and naturalness, stars indicate significant results (* indicates $p < 0.005$, ** indicates $p < 0.001$). Stimulus labels are explained in Table 9.3.

syllabic pulsation, no significant difference is found, although the rating of S+ – is somewhat lower. All stimuli with at least one 'inactive' feature (the dashes in S+ –, S– +, S– –) do not differ significantly from each other. This indicates that, in order to induce a difference in the degree of perceived amusement, one would need more than one feature to be 'active'.

The results to the question whether the features influence the degree of perceived *naturalness* are shown on the right side of Figure 9.5. A Friedman test (same 23 subjects, χ^2 = 60.91, $p < 0.0001$) reveals two subgroups, splitting the set of stimuli into a 'human' and a 'synthetic' category. Post-hoc testing (as described above) only indicates that the human stimulus differs from all synthetic ones. The slightly more natural ratings of the stimuli *without* syllabic pulsation match a comment made by a subject, namely that the stimuli *with* pulsation (S++, S+ –) sounded rather unnatural.

Furthermore, we analyze whether the results show a correlation between the stimulus preferences over the two categories *amusement* and *naturalness*. A non-parametric correlation test (Spearman-Rho) reveals a correlation coefficient of 0.86 ($p < 0.001$, two-tailed). This would mean that the stimulus that is selected as showing a higher degree of amusement is very often also selected as the more natural sounding stimulus. This strong correlation is, however, largely due to the high ratings of the natural stimulus in both perception categories.

Finally, the absolute ratings of naturalness of the "more natural" stimulus of each pair are analyzed by conducting a one-way ANOVA (level of significance 5 %). The results are shown in Figure 9.6. They show an overall significant difference in ratings ($p < 0.001$, $Z = 201.11$). Post-hoc testing following the Scheffé procedure indicates that only the original stimulus (H++) versus the synthetic stimuli differs significantly in the ratings of naturalness. Thus overall, the subjects again make a clear distinction between human and synthetic stimuli.

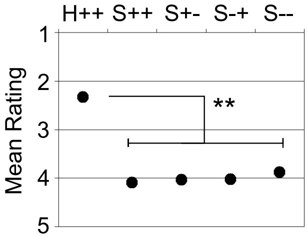


Figure 9.6: Mean absolute speech-laugh ratings of naturalness on a scale from 1 "very natural" to 5 "really unnatural." Only H++ differs from other stimuli in a significant way. Stimulus labels are explained in Table 9.3.

9.4 Discussion

Based upon descriptions of speech-laugh phenomena, we presented hypotheses about their production mechanisms and transferred them to control strategies for articulatory speech synthesis. We imitated a short speech-laugh from a corpus of spontaneous speech and manipulated the two parameters *syllabic pulsation* and *smiled laugh-vowel quality* independently of each other. The different speech-laugh versions were evaluated for perceived amusement and naturalness. To conclude, we first discuss phonetic aspects regarding this experiment, followed by synthesis technical aspects.

9.4.1 Phonetic aspects

First, we discuss the main results of the perception experiment, which may have set the stage to investigate additional speech-laugh features in future work, which is briefly sketched. We then point to a few design issues that are connected to the phonetic evaluation of the experiment. Finally, we extend our view for a final conclusion on the phonetic aspects of speech-laughs, laughs, and smiled speech.

To sum up the results regarding perceived amusement, evidence is found that an activation of *both* speech-laugh features yields higher amusement ratings than when only one or no features are active. This can be interpreted as a basic confirmation of the implementation strategy regarding the speech-laugh features. Furthermore, our data suggest that the subjects perceive the speech-laughs featuring *syllabic pulsation* as coming from a more amused speaker than speech-laughs that do not feature syllabic pulsation. A significant difference in amusement perception induced by the presence or absence of a *smiled laugh-vowel quality* has not been found. However, we assume that the acoustic difference of the two vowel qualities was too subtle in our experiment, so there is no reason to exclude smiled vowel quality from the complex system of speech-laughing.

The amusement rating depends more on rhythmic than on spectral characteristics. The influence of syllabic pulsation might indicate a confirmation of observations found by

Nwokah et al. (1999), who report that this kind of rhythmicity is a very salient feature of speech-laughs.

Besides rhythmic information, overall duration may also play a role in amusement or mood perception. Vettin (2003: 21–23) reports that long laugh bouts (i.e. laugh sound sequences within one exhalation phase) cause a higher rating of perceived laugh frequency than many short laugh bouts. This is correlated with a better mood that a listener ascribes to the speaker. In comparison, we investigated speech-laughs – not laughs –, taken out of context, and they were all of equal duration (1.1 s). Despite the equal duration, they received different amusement ratings. This seems to indicate that the duration of a speech-laugh in terms of seconds does not correlate with mood perception. On the other hand, if we count the syllables as an indication of perceived length, our results agree with those of Vettin (2003): More syllables produce higher amusement ratings. However, the number in our stimuli ranges between one and two elements, while Vettin (2003: 23) refers to "long" bouts as having more than three elements.

Future work on speech-laughs could study the impact of the additional phenomena described in Section 9.1.2, such as inserted frictions, vocal tremor or vibrato, temporal structure, and combinations of these features. Preliminary synthesis trials indicate that, in principle, these phenomena can be imitated with the synthesizer. The impact of more distinct smiled-vowel qualities in speech-laughs should also be investigated, e.g. by creating different settings for the lips and possibly the tongue.

Three phonetic aspects, which are also intertwined with the evaluation test design, are discussed next. Firstly, since speech-laughs can originate from different emotions, one might want to evaluate other dimensions than 'amusement' in a perception test. For this speech-laugh, it was clear from the context in the dialog that it is indeed displaying positive emotion, being accompanied by a laughing sequence. However, other emotions such as fear or shame are imaginable as well.

The second aspect concerns an additional evaluation dimension. Particularly with speech-laughs, issues of speech intelligibility should be addressed in future work. It could be argued that laughed speech reduces the intelligibility of the content but at the same time increases affective clarity.

The third aspect concerns the design of the evaluation scale in the last subtask of the experiment. Absolute naturalness was evaluated using a 5-point scale to give the participants a free choice of degree of naturalness to gain insight into the overall synthesis quality in terms of naturalness. However, future work should also consider floor and ceiling effects, and perhaps a scale with more points should be used.

To conclude the phonetic considerations regarding speech-laughs, we extend our view to encompass all three laughter-related phenomena that have been the subject of our experiments: laughed speech, laughs (Chapter 8), and smiled speech (Chapter 6). One aspect not mentioned in any of the evaluation schemata so far is the multi-modal nature of these phenomena. We have only studied the acoustic (vocal) channel in our experiments. The information and influence stemming from the visual component of laughter during the audio perception tests has not been considered at all, although this seems historically the primary component of laughter. Especially for smiled speech, and 'smileyness' in laughed-speech, it would be of interest to find out how the modality influences perception.

Distinguishing between laughs, speech-laughs, and smiled speech from an articulatory point of view, we can summarize that laughs place an increased demand on glottal control, and not so much on the supraglottal mechanism. Speech-laughs, on the other hand, combine the segmentally induced supraglottal demands of speech with the glottal and subglottal demands of laughter. Lastly, smiled speech combines the 'regular' speech mechanisms with competing supraglottal demands regarding segmental targets of 'smileyness'.

We still have relatively little knowledge about how exactly these different mechanisms are used to enrich a theoretical, 'neutral' baseline of speech. Thus, a general question in future work could be whether it suffices to simply superimpose laughter features onto 'neutral' speech by subsequently adding single manipulations. Our hypothesis is that simple superposition would not work since acoustic analyses have shown coordinative adaptation between the laughter system and the speech system. While corpus analyses present important groundwork because we can capture the diversity of the phenomena only by analyzing real-life, spontaneous instances, a subsequent imitation with articulatorily transparent speech synthesis, as in our experiments, perhaps helps to better understand the mechanics of laughs, speech-laughs, and smiled speech.

9.4.2 Synthesis technical aspects

Overall, this experiment represents a first investigation on articulatorily synthesized speech-laughs. It was guided by corpus observations on how speech and laughing might interact. Regarding the technical aspects of the experiment, we found that VTL is indeed capable of imitating phenomena as diverse and complex as this speech-laugh event. The gestural score proved in principle to be a feasible tool to construct such 'hybrid' utterances that account both for linguistically induced supraglottal *speech* activities and the glottal and subglottal *laugh* activities. At the same time, the gestural score is a way of visualizing the interaction of these two dynamic systems (laughing and speaking, cf. Nwokah et al., 1999) merged into one outcome (laughed speech).

Compared to the imitation of a laugh (Chapter 8), we have to account for both laugh elements and segmental intelligibility. On the one hand, this represents a complex task but on the other hand it seems that VTL can handle the speech-laugh imitation more easily than the laugh imitation. Firstly, to cope with the complexity of the speech-laugh, it is very useful that the parameters in VTL are not abstract or acoustically driven, but clearly defined in the articulatory domain. Therefore they are directly accessible for manipulation, enabling utterances that differ greatly from 'regular' segmental quality. Secondly, in contrast to the laugh imitation, VTL did not face any technical limits of the synthesizer during the speech-laugh imitation. Extreme values such as high lung pressure levels probably occur with laughs rather than with speech-laughs because the latter are somewhat closer to the demands of ‘regular’ speech. Finally, the imitation of speech-laughs is easier in VTL than the imitation of laughs, perhaps because VTL has been designed primarily as a *speech* synthesizer.

However, there is currently a limitation in the technical assessment. It is due to the unknown influence of two different aspects of the acoustic performance of the synthesis system on perception. On the one hand, the low ratings in the naturalness question can be viewed as an illustration of how difficult it is to articulatorily imitate a speech-laugh with the chosen articulatory strategy in particular. So perhaps the perceived naturalness is rated so low because the human speech-laugh elements need to be imitated with additional, or perhaps different articulatory strategies.[2] On the other hand, the ratings might be negatively influenced simply by the synthetic nature of the acoustic simulation. It is an open question whether the low ratings stem from the synthetic timbre of the acoustic simulation or whether the articulatory strategies need to be refined. Presumably, it is a mixture of both.

These considerations might also interact with a basic problem in laughter evaluation. As mentioned in the discussion of the laugh experiment (Section 8.4), the evaluation of the speech-laughs seems to be a harder task for naive listeners than the evaluation of ‘regular’ speech. Perhaps as a result, the ratings are low, and even the natural stimulus does not receive the highest possible score. With ‘regular’ speech, listeners tend to use their expectations of how a word should sound in order to rate the word. Expectations for ‘good’ sounding speech-laughs or laughs may not be so clear-cut because there might not be an established system of good candidates behind the diversity of elements in speech-laughs and laughs. Laughing operates in a paralinguistic domain which runs in parallel to the linguistic content and might feel out of place in general when pushed into the foreground. If this were the case, it could also have contributed to the low overall naturalness ratings.

Despite all the challenges, it would be a great benefit for synthetic voices to incorporate these paralinguistic phenomena in their speech because the appropriate usage of these affective vocal signals can make synthetic utterances sound more naturalistic.

[2] One participant e.g. noted regarding a synthetic stimulus at the beginning of the test that it “sounded like a goat.”

Chapter 10

Experiment VI – Articulatory setting of Saxon-accented vowels

Similar to the information of age in the voice (Chapter 7), speakers often also reveal their regional origin in the way they speak. The regional sound may include specific segmental properties as well as typical behavior in voice quality, intonation, and rhythm. How exactly the typical sound is achieved, is subject of ongoing research, which has a long-standing history. The aim of this experiment is to contribute some articulatory insights.

Usually, an individual's regional accent has developed over years and belongs to a speaker as a personal trait. Besides being associated with a specific speaker it can also convey certain stereotypes of a whole group of speakers. Just as with age information, incorporating regional accent information is an aspect that can make synthetic voices sound more natural and more individual.

In this chapter, we present a comparative investigation of articulatory configurations of six high and mid-high long vowels, extracted from two-syllable words produced in reading style. They belong to two different German accents, Saxon (*Sax*) and Standard High German (*SHG*). We use the term 'accent' as a shorthand for accent-colored speech, denoting a 'colored' pronunciation as opposed to a strongly dialectal pronunciation.

The experiment is motivated by the notion of *articulatory setting* (AS), or *Artikulationsbasis*, and the aim is to find out whether the two accents show overall systematic shifts in the positioning of the articulators. We employ a special acoustic-to-articulatory inversion algorithm within VocalTractLab to create the articulatory vowel simulations. They are evaluated by an informal listening test, visual inspection of the articulation, formant measurements, and a formal perception test (for the latter, see Chapter 11).

The speech production goal of this experiment is to test the hypothesis that different accents of German are realized by *systematic* shifts in the articulators.

The technical evaluation goal of this experiment involves two aspects. The first one deals with assessing the functionality of a newly implemented optimization algorithm that helps to adapt articulation to match given acoustic data. The second is to assess whether VocalTractLab is suitable for synthesizing differences in articulation as small as those between standard and regional varieties of High German, i.e. pronunciation varieties that diverge only slightly from the Standard pronunciation.

In terms of the overview of all experiments, this chapter presents an illustration of exclusively supraglottal manipulations, focusing on the movements of the lips, jaw and tongue.

In Section 10.1, we provide background information on regional accents in general, complemented by a brief description of typical characteristics of Saxon and relevant features for resynthesis. In Section 10.2, we explain the creation of the stimuli and describe the evaluation procedures of Saxon vs. Standard High German vowels. The results are presented in Section 10.3, followed by a discussion.

This is the first of two chapters dealing with pronunciation details in these two varieties. Based on the vowels created in the current experiment, the imitation and formal perceptual evaluation of *words* spoken in the two varieties is presented in Chapter 11.

10.1 Aspects of accent perception and production

This section provides background information on the concept of *Artikulationsbasis* or articulatory setting and its definition(s) (Section 10.1.1), followed by an introduction of how it has been investigated and what particular challenges are linked to this research (Section 10.1.2). These general considerations are then complemented by a description of typical features of Saxon (Section 10.1.3), from which the synthesis features for VTL are derived (Section 10.1.4).

A native listener of a language can easily perceive if someone speaks with an ‘accent’ that does not represent the standard pronunciation of the language, or indeed if it deviates from their own non-standard accent. This is true for foreign-accented speech as well as for regional-accented speech. Studies have shown that both phoneticians and naive listeners are able to discriminate native from non-native speech (Flege, 1984; Flege and Hammond, 1982; Magen, 1998).

Each language or regional or social variety thus has a typical overall “sound” (e.g. Gick et al., 2004: 220, Schaeffler et al., 2008) or “idiomatic phonetic character” (Hammarberg and Hammarberg, 2009: 75), which can serve as a “marker” of that community (Mennen et al., 2010: 14). It presumably is created by articulatory characteristics that differ across varieties and languages. However, it is difficult to capture what exactly contributes to this typical sound.

The above considerations may be interpreted as referring to the notion of articulatory setting (AS) or *Artikulationsbasis*. While this concept has been around for centuries,[1] only recently could articulatory data be presented that claim to support its existence (Wilson, 2006, discussed further below).

[1] For an historic overview and evolution of the concept, see also Laver (1978).

Findings and implications from AS research have been used e.g. to automatically detect regional accent in English, based on relative formant differences (Barry et al., 1989), and to calculate formant frequency distance metrics for accent recognition and for clustering into different groups (ACCDIST, Huckvale, 2004, 2007a,b). A better understanding of articulatory settings can also support the development of regional and foreign accented synthetic voices.

10.1.1 Concept of articulatory setting

The general concept of a 'basis of articulation' (Sweet, 1890a: 69) has been addressed under different names and with slightly different definitions, some of which encompass the position and movement of the articulators, while others explicitly subsume under AS more levels than only supraglottal articulation.

Beatrice Honikman (Honikman, 1964) sums up the phenomenon as "the gross oral posture and mechanics" (p. 73), and she introduces the term *articulatory setting*. It is used in other studies as well, with slightly differing definitions, e.g. by Wilson and Gick (2006: 150): "the underlying setting of the tongue and other articulators during speech"; Wilson (2006: ii): "a language's underlying or default posture of the articulators"; Ramanarayanan et al. (2010: 1994): "the gross articulatory posture deployed as the default basis from which economic and fluent production a (sic!) language occurs." Gick et al. (2004: 220) also speak of a "postural basis."

An explicitly more comprehensive definition, which includes voice quality aspects, can be found e.g. in Hammarberg (2009: 13), where "articulatory settings (Artikulationsbasis)" refer to "the language-specific positions and gestures of the voice and articulatory organs in speech." (See also Hammarberg and Hammarberg, 2009.) Similarly, Mennen et al. (2010: 13) state that 'phonetic setting' means a "tendency to make the vocal apparatus employ a language-specific habitual configuration," giving the example of "degree of lip-rounding, tension of the lips and tongue, jaw position, phonation types, pitch range and register." Following Laver (1994: 399), the concept of articulatory setting "is applicable at every level of phonetic analysis", including the articulatory, the phonatory and the prosodic level as well the level of overall muscle tension. Mennen et al. (2010: 14) also report the additional terms "voice-setting features" and "voice-quality settings." A closely related but more restricted term is suggested by van Buuren (1995: 130), using "postura" as a "phonetic term for the relevant vowel-type and approximant-type postures of tongue and lips."

In the following we use the term *articulatory setting* to refer to language-specific positions and gestures of the whole vocal apparatus. However, in our experiment, we focus on supraglottal language-specific manipulations.

10.1.2 Techniques in AS research

Regardless of the name of the concept, it was traditionally investigated by self-observation, since appropriate tools were not available for detailed empirical investigations (cf. also Chapter 2). So for a long time articulatory settings were described by means of impressionistic accounts, based on analytic listening, reproduction and verbal description (see e.g. Honikman, 1964). Later, acoustic methods were also used, inferring single articulation properties from the acoustics. In more recent times, articulatory methods have contributed to AS research. An additional approach can be taken by inverse acoustic-to-articulatory mapping techniques.

We discuss acoustic and articulatory methods, their challenges with articulatory settings and suggest how to apply acoustic-to-articulatory inversion to AS research in our experiment.

Acoustic techniques

Mennen et al. (2010) discuss pitch range measures, formant frequencies, and long-term average spectra (LTAS) in relation to the investigation of articulatory settings. Formant frequency measurements seem the most suitable method of those three, and "might also be able to provide information about language-specific phonetic settings" (Mennen et al., 2010: 29). They represent a "rather fine-grained approach of formant analysis" which "might primarily be adequate for the analysis of supra-laryngeal settings" (Mennen et al., 2010: 33). Since articulatory settings are learnt by speakers, they must be "extractable from the acoustic signal" (Mennen et al., 2010: 35).

Based on known general relations between vocal tract shape and F_1-F_2 behavior, acoustic findings can cautiously be interpreted articulatorily, as suggested for instance by Mennen et al. (2010: 29) : "A more open jaw setting should result in a general raising of the first formant (F_1) across different vowel categories, and a relatively fronted tongue body position should result in a general raising of the second formant (F_2)." Studies involving formant measurements have been carried out e.g. by linking formants to phonetic settings of different languages (Lowie and Bultena, 2007), or by synthesizing sounds from systematically changing vocal tract shapes (Story and Titze, 2002), detecting formant patterns throughout the signal, which according to Mennen et al. (2010: 30) affirm "the validity of formant analysis when investigating phonetic settings." Cf. also Reetz (2003: 137ff), describing the general relations between vocal tract cavity dimensions and formant frequencies.

Articulatory techniques

However, as Mennen et al. (2010: 17) point out, acoustic techniques have "the problem of ambiguous mapping that occurs when aspects of articulation are inferred from the acoustic domain." In contrast, "articulatory techniques have the advantage of providing direct access to vocal tract configurations" (see also Chapter 2). Mennen et al. (2010) discuss flesh-point tracking techniques, ultrasound tongue imaging (UTI), and electropalatography with respect to phonetic settings. Ramanarayanan et al. (2010) employ "automatically-extracted features to quantify areas of different regions of the vocal tract as well as the angle of the jaw." (p. 1994) They collect their articulatory data with real-time MRI.

While both techniques have their strengths and typical problems, research in AS shows two general challenges. First, it is hard to reliably separate the underlying articulatory setting from the segmental or phonemic demands on the vocal system during speech. This may easily lead to a 'phonemic confound' in articulatory setting investigations (cf. Mennen et al., 2010: 15). AS research has therefore been linked to investigation of so-called interspeech postures (ISPs, Gick et al., 2004), also mentioned as "intermediate 'speech posture' " (Perkell, 1969: 52). ISPs are understood as rest positions between utterances, when the vocal tract is ready to speak but has not yet started speaking – as opposed to absolute rest positions, where no speaking is planned. Wilson (2006) used ISP "as a measure of AS" (p. ii), claiming that his research "empirically confirms centuries of non-instrumental evidence for the existence of AS" (p. ii). As Schaeffler et al. (2008: 121) point out, the advantage of ISPs is that one can collect data "without confounding effects from lexis, phonotactics or phonological inventory." Mennen et al. (2010: 34) note, however, that "without further evidence we should not be seduced into thinking that the postures measured during such pauses are necessarily the same as traditional conceptions of phonetic settings that are based on impressions of speech."

The second general challenge in AS research involves the issue of normalization (Mennen et al., 2010), which is a typical problem of studies involving more than one speaker, due to differences in physiology.

Inversion experiment

The acoustic and articulatory methods discussed above are all based on recorded speech data, acoustic and articulatory. In contrast, with the method of articulatory inversion, one can generate tailored data based on acoustic targets, and use them for further articulatory analyses. General considerations about inversion have been introduced in Section 2.2, in Section 4.2.4 we presented the particular inversion algorithm implemented in VTL.

In this thesis, we employ restricted-scenario acoustic-to-articulatory inversion to find out whether vowels of two German accents show systematic supraglottal articulatory shifts. Our experiment is motivated by the idea of articulatory settings (AS), focusing on the supraglottal aspects of this concept. To reduce the phoneme-confound issue, we compare only phonemically equivalent vowels (pairwise), and analyze whether the intra-vowel-pair comparisons point in a particular direction for all vowel pairs. To avoid problems of normalization, we use speech material for both varieties stemming from one bilingual ('bi-dialectal') speaker.

In the following, we discuss articulatory descriptions of Saxon vowel production (Section 10.1.3), and deduce from that the articulatory variables for our simulation experiment as well as articulatory hypotheses (Section 10.1.4).

10.1.3 Characteristics of Saxon vowel pronunciation

The Saxon accent of German (Upper Saxon German, Obersächsisch) is spoken in the region of Dresden, Leipzig and Chemnitz, and belongs to the East Central German (ECG) dialect group (Ostmitteldeutscher Dialektraum). It is clearly perceivable by laymen and therefore seems to be an obvious German accent. This implies that there should be pronunciation differences between Standard High German and Saxon. So some of them might be reproducible in our experiment. The variety that we analyze is not a strong dialectal pronunciation (*Basisdialekt*) but a more standard-like, colloquial style regional accent pronunciation, i.e. a Saxon-colored regional variety of Standard German pronunciation. For an orientation on the standard–dialect continuum see e.g. Lameli (2004: Chapter 2). The speaking style that we analyze is word-list reading style.

Thus, we compare reading-style Upper Saxon German (*Sax*) against reading-style Standard High German (*SHG*). The latter we regard in this experiment as the 'unmarked' pronunciation, the Saxon variety constitutes the regionally marked pronunciation. In this sense, we report findings in the form of divergences from the unmarked Standard German setting.

We only analyze vowels, more specifically a part of the vowel set. However, analyses of the other vowels and consonants are desirable. The vowel set in our analysis comprises the high and mid-high long vowels /uː oː yː øː iː eː/.

We follow prevalent impressionistic accounts that describe typical Saxon pronunciation as 'involving a protruded lip setting and a certain movement of the jaw', to determine which articulators we should put the focus on. The reason why we focus on a subset of articulators is because an unrestrained approach would yield too many degrees of freedom for the inversion mapping task. As we will see however, the focused articulators account for much of the variation in the vocal tract space anyway.

Rues et al. (2007: 93ff) give phonetic indications on relevant features of pronunciation by summarizing reports in the literature. We list those features which are connected to our stimulus set of high and mid-high vowels. Main differences to Standard High German pronunciation are thus:

- Derounding of the vowels [øː yː] into [eː iː]
- Lowering of vowels before <r>, e.g. [eː] to [ɛː]
- Monophthongization of [a͜ɪ a͜ʊ] (Middle High German <ei>, <ou>) into [eː oː]
- Diphthongization of [oː] into [o͜ʊ]

In this list, the derounding and lowering tendencies are of primary interest for our experiment. The reported processes of monophthongization and diphthongization will not be considered further because they involve temporal features regarding transitions from the first to the second part of a diphthong. Since our experiment is a first attempt to simulate fine articulatory differences in regional pronunciation, we focus on static features such as vowel target configurations in the vocal tract. As will be brought up in the discussion again, dynamic, or temporal features, should be investigated as well to complement this line of research.

10.1.4 Transfer into synthesis features and articulatory hypotheses

Based on the articulatory differences identified above, the main features that are manipulated in VocalTractLab to create Saxon vowels are the supraglottal parameters. However, it is noted that e.g. voice quality features or prosodic elements probably also have an impact on the typical 'sound' of Saxon, as this is reported for other German regional accents (Braun, 1988, 1996, 2001; Braun and Wagner, 2002).

The variables under investigation can be grouped into parameters of the lip, the tongue, and the jaw. The velum is excluded from a main-influence role because we assume for our imitation that all vowels should have the same oral quality.

Since the vocal tract in VTL is represented by a geometric model, we are prepared, again (as before, e.g. in Chapter 5), that certain co-dependencies prevailing in natural human articulatory settings may be missing. One example in our experiment here is that the tongue and the jaw show no explicit co-movements.

We hypothesize that the Saxon accent has a lowered and fronted tongue-body position for the long high and mid-high vowels /uː oː yː øː iː eː/. An articulatory limit to the fronting shift is expected for the long front vowels /yː øː iː eː/. However, the rounded vowels may become 'derounded' to compensate this. The term 'derounded' is used to refer to vowels that, in a reference system, are supposed to be *rounded*, i.e. Standard High German rounded vowels, but are not pronounced with the common rounded-lip setting by the speakers. Instead, they may be spread vowels or slightly derounded vowels, i.e. articulated with less lip rounding and protrusion than common for the standard rounded vowels.

10.2 Data and analysis

In this section, we describe the recording and acoustic characterization of the speech material in the two accents Saxon and Standard High German (Section 10.2.1). This material constitutes the basis for the subsequent articulatory simulations in VTL (Section 10.2.2). The simulated sounds are subjected to different analyses, described in Section 10.2.3.

10.2.1 Human speech data

Recordings

To build a speech corpus with relevant target words, we design a wordlist covering the vowels /uː oː yː øː iː eː/ as stressed target vowels in different phonetic contexts. We ask an adult male speaker to read the wordlist in different accents: *i)* using Saxon-accented reading style (Saxon-accented read speech, *Sax*), and *ii)* using Standard High German reading style (Standard High German read speech, *SHG*). In Section 10.1.2 (p. 149), we touched on issues of normalization of the vocal tract between speakers. To avoid this problem we record a speaker who speaks both Saxon and Standard High German as mother tongues. Prior to the recording of the Saxon reading style, the speaker immerses himself in the target variety by listening to authentic Saxon speech recordings.

Our recordings are carried out in a sound-treated room with a digital USB-microphone and a laptop computer, using the recording feature of the phonetic software Praat (Boersma, 2001). The list of recorded words can be found in Appendix Tables F.1 and F.2.

Acoustic analysis procedure and results

All regional carrier words of a given target vowel are informally evaluated auditorily by the author with regard to the strength of their accent realization (strong/weak/very weak), and the strong candidates are subjected to formant analysis. In each 'strongly' accented word we annotate the stationary part of the target vowel with one of the labels /uː oː yː øː iː eː/. Using the software Praat (Boersma, 2001), we calculate the first three formant (F_1, F_2, F_3) of these segments, using standard settings to analyze male voices, in which the formant tracking algorithm allocates five formants in a frequency band between 0 and 5 kHz with a window length of 0.02 s.

Only the good-quality recordings are used for further analysis, and vowel token pairs are selected from the same carrier word in both accents so as to minimize context effects on formant frequencies and to have the same wording available for later use. 'Good' recording quality denotes modal voice quality, not too much coarticulation or reductions, and overall good acoustic recording quality so that the formants can be extracted without any problems.[2]

One representative token for each target vowel is selected to serve as the 'model' for the imitation with VocalTractLab. The imitation procedure is described in Section 10.2.2. The mean formant frequencies of the stationary parts of the selected representative vowels are depicted in the F_1-F_2 plot in Figure 10.1a. An additional plot, including F_3 values, is presented in Figure 10.2a. The mean formant frequencies can be found in Table 10.2 (p. 159) in the rows labeled *Target*. A comparison with the synthesized vowels is discussed on page 161.

The acoustic shifts from Standard High German to Saxon-accented German can be interpreted articulatorily as follows: An increase of F_1 assumes a more open manner of articulation, e.g. by a lowering of articulators. This applies to all sounds but /oː/. An increase of F_2 suggests a more anterior articulation, which can also be induced by lip spreading, or less protrusion of the lips, which also reduces the front cavity in the mouth. This applies to all rounded vowels, but not to /iː eː/ which are front and spread already. Lip rounding, i.e. protrusion and constriction of the lips, usually lowers all three first formants, especially F_3 and F_2 (Neppert, 1999: 132f, Stevens, 1998: Chapter 6). In all rounded vowels, we find an increase of F_2 and F_3 in Saxon, the only exception being F_3 in /oː/. Articulatorily, this may indicate a derounding and/or fronting tendency, showing less protrusion and constriction at the lips.

[2] Due to recording circumstances some words were recorded in less good quality. So this pairing of identical words was achieved in all but one instance, /uː/, cf. also Table 10.2, p. 159.

In Section 10.2.2, we describe how we re-synthesize the natural vowels that have been analyzed above. The synthetic vowels will be analyzed acoustically in the same manner as above, for comparison with their natural counterparts. Additionally, they will be analyzed articulatorily to find systematic shifts in articulation between the two accents. The analysis procedures are described in Section 10.2.3. Results are presented in Section 10.3.

10.2.2 Synthetic speech data

The aim is to create two sets of imitations of the vowels analyzed in Section 10.2.1: From the Saxon-accented vowel set we create synthetic imitations that resemble the acoustic structure of the natural Saxon vowels (intended Saxon accent), and likewise we create a second set that resembles the acoustic structures of the natural Standard High German vowels (intended Standard German accent).

The optimization algorithm introduced in Section 4.2.4 is applied to the default settings of VTL. It adjusts the default vocal tract configurations in such a way that the vowel formants match those of the natural Saxon ones. Based on the target values of the first three formant frequencies of each target vowel, we thus create a new vowel set that resembles intended Saxon vowels (from reading style with Saxon accent). In the same way, we simulate the Standard High German pronunciation of the recorded natural vowels. All in all, the simulation produces twelve new phone definitions: /uː oː yː øː iː eː/$_{Sax}$ and /uː oː yː øː iː eː/$_{SHG}$.

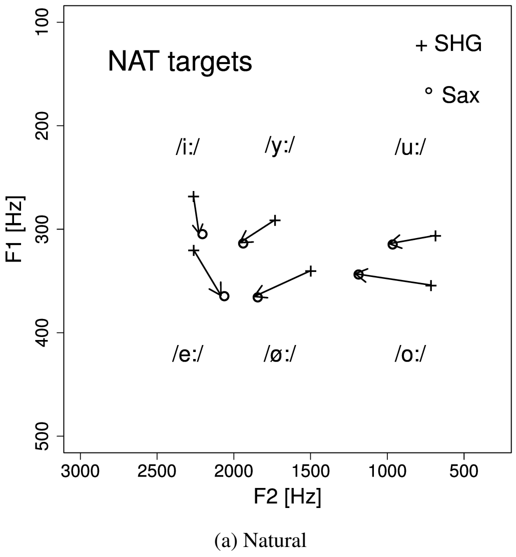
For the optimization tasks we assign full weight to the parameters that were identified in Section 10.1.4, i.e. lip, tongue and jaw, so they have a large influence on the probabilistic stepwise changes of the vocal tract configuration during execution of the optimization algorithm and will produce the main articulatory shifts. Other VT parameters are set to a weight of 10 %.

The optimization process itself is executed following the general procedure described in Section 4.2.4 (p. 56). The target frequencies are input into the optimization algorithm vowel by vowel. The values are documented in Table 10.2, in the *Target* row of each vowel, based on the formant analysis presented in Section 10.2.1. With a step size of 1 % or 2 %, which seems to work best for this task (cf. Section 4.2.4, p. 57), we execute several runs of the algorithm for each vowel (one run consists of 100 iterations), and document the formant frequencies finally attained and their corresponding vocal tract configuration for later use: in the speaker definition file for later application in synthesis and for acoustic inspection, and as vector graphics for later visual comparison.

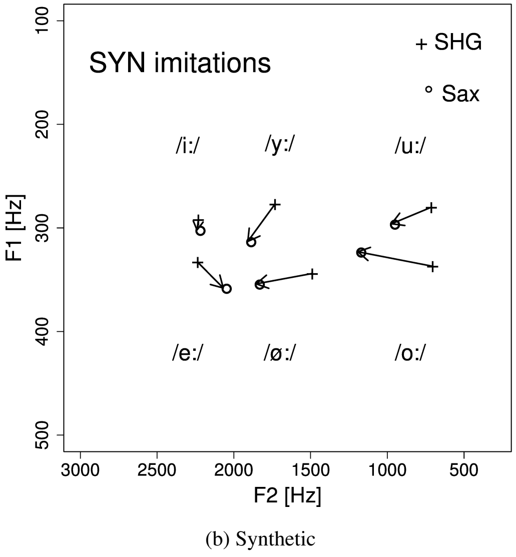
In the end, we have obtained several articulatory suggestions for each target vowel in each of the two accents.

10.2.3 Analysis methods

The analysis of the newly created vowel configurations is two-fold. The first part is carried out in the acoustic domain by informal listening tests, narrow phonetic transcription and

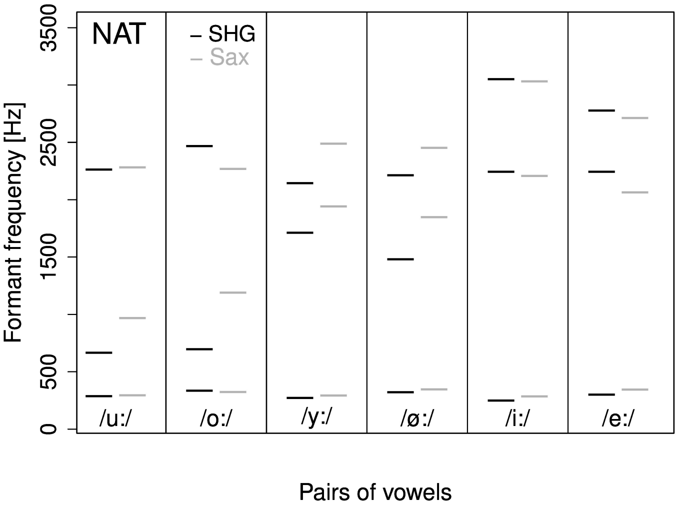


(a) Natural

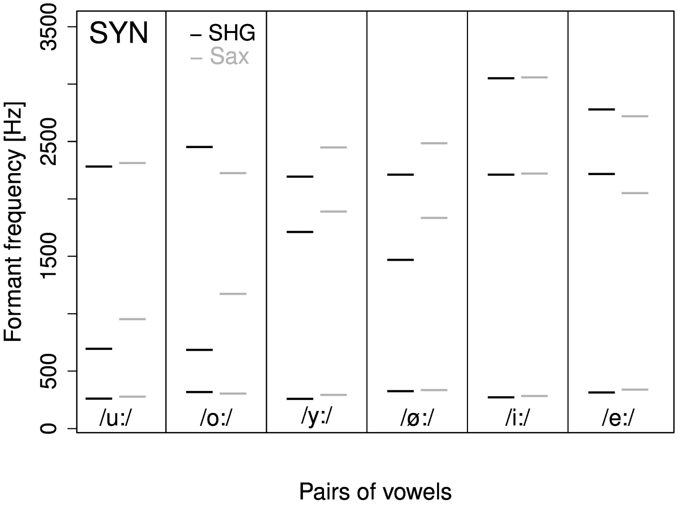


(b) Synthetic

Figure 10.1: Formant frequency plots of the isolated vowels: (a) of the natural speaker, (b) the imitations created with the optimization algorithm, for each accent. NAT: natural, SYN: synthesized, SHG: Standard High German, Sax: Saxon-accented German. The arrows connect a Standard High German vowel with its Saxon-accented counterpart to point out the direction of acoustic shift.



(a) Natural



(b) Synthetic

Figure 10.2: Plots depicting (a) the first three formants of the natural vowels used as imitation targets, and (b) the synthesized vowels. NAT: natural voice, SYN: synthetic voice, *SHG*: Standard High German, *Sax*: Saxon-accented German. Lowest bar in each column = F_1, middle bar = F_2, top bar = F_3.

formant analysis. The second part of the analysis is located in the articulatory domain, where we compare the vocal tract shapes visually. The acoustic evaluation is carried out first because the optimization task was successful from a phonetic point of view only if the (accent) differences are indeed perceivable by the ear.

Acoustic-auditory evaluation procedures

One 'top-candidate' articulatory suggestion of each vowel was selected to be subjected to three acoustic-auditory inspections. The 'top candidate' was selected on the basis of it sounding very similar to its natural counterpart and of it being of a generally good sound quality. Other candidates, which the optimization algorithm produced, sounded rather muffled and were not subjected to further analyses.

Firstly, an informal auditory assessment evaluated whether the regional vowels sounded different than their Standard High German counterparts. To this end, we simulated a simplified word context by synthesizing a set of logatoms /dVdVdV/, with $V = /\text{iː eː uː oː yː øː}/$, containing either a *Saxon* or a *Standard High German* vowel. All other gestures were kept constant in the articulatory synthesis system, so the vowel target was the only differing feature, e.g. $/\text{duː}_{Sax}\text{duː}_{Sax}\text{duː}_{Sax}/$ vs. $/\text{duː}_{SHG}\text{duː}_{SHG}\text{duː}_{SHG}/$.

Secondly, the vowels were evaluated auditorily in isolation, namely by narrow transcription by phonetically trained listeners. They listened to a vowel as often as they wanted to, via a set of headphones which was the same for all transcribers to ensure comparable acoustic output. The trained listeners were asked to transcribe the vowel quality with special emphasis on diacritics to determine the perceived place of articulation as precisely as possible. For orientation, we equipped each listener with a visual transcription guideline, consisting of the IPA vowel chart and a set of diacritics that describe deviations from the base phoneme regarding vowel height, lip rounding, and horizontal tongue body position.

Thirdly, a formant frequency plot was calculated to see whether acoustic effects were as expected. F_1 to F_3 were again measured with Praat (Boersma, 2001), as described in Section 10.2.1.

Articulatory evaluation procedure

We created multi-layer figures (cf. Figure 10.3, p. 160) by stacking up the graphs of the vocal tract shapes of the most suitable candidates of the output of the optimization algorithm (obtained as presented in Section 10.2.2). Suitable candidates were solutions that had a low formant error (definition see 4.2.4) and an appropriate auditory quality, i.e. closely resembling the target vowel quality. The top candidates of the acoustic evaluation belonged to this group. The multi-layer figures allow for direct visual comparison of articulatory settings. We evaluate articulatory differences a) across several suitable solutions for a vowel within one regional variety, and b) across the two accents. We compare the visual interpretation to our initial hypotheses about articulation shifts.

Table 10.1: IPA transcriptions of the two vowel sets obtained by the optimization algorithm introduced in Section 4.2.4 to imitate Standard High German pronunciation (*SHG*) and Saxon accented pronunciation (*Sax*), synthesized using FDS mode (cf. Section 10.3.1, p. 158). The transcriptions are provided by six trained phoneticians, slashes indicate multiple suggestions by one person. Diacritics used (from top to bottom): centralized [e̽], nasalized [ẽ], more rounded [e̹], less rounded [e̜], retracted [e̠], advanced [e̟], raised [e̝], lowered [e̞].

Vowel	SHG	Sax	Vowel	SHG	Sax
/iː/	i̠ i̞ ɪ i i i̽ i̞/e̝	ɪ̞/e̠̝ e̝ e i̞ e̝ e̝	/eː/	e e e̠ e e/e̝ e	e e e̠ e̝/e̝̽ e̹̠̞ e̞/e
/uː/	ʊ u̞/u ʊ u u̽ u̟	u̜̟ ʊ̜ ɯ̹ u̟̝/u̟ ɵ u̜/ɯ	/oː/	o/o̝ o o o o̞/o̟̞/o̽ o	ɵ ɤ ø̞ ø̝ ø̠ ø̽
/yː/	ʏ/ʉ̞ y ʏ y y̽ y̜/y̽	ø̜̠/ʏ̜̞ y̽/ø̽ e̹̽ y̜̞ y̽ e̹	/øː/	ø ø ø̠ ø ø ø	e̹/ø̜ ø̜/e̹ ɘ ø̜̞ ɵ̟/e̹̠ e̹̞

10.3 Results

In the following, we present the evaluation results of the six synthesized Saxon and Standard High German vowel pairs, created by the formant optimization algorithm, based on the formant structures extracted from natural carrier words. The vocal tract parameters of the best simulation for each accented vowel are provided in Appendix Table F.4 for reference.

10.3.1 Acoustic-auditory evaluation

Informal listening

The informal listening evaluation based on the [dVdVdV] logatoms suggests a clear effect of accent variety, which is most pronounced in rounded back vowels and least perceivable in unrounded front vowels. Thus, it seems that the optimization algorithm worked successfully in general, and in the following, the vowels are evaluated more thoroughly.

Transcriptions

The results of the narrow phonetic transcription are presented in Table 10.1. Main perceived differences between the synthesized accents are a fronting and derounding tendency, especially in the rounded vowels and back vowels.

The transcription task reveals two problems with the synthesized vowel tokens. The first problem concerns the vowel /uː/, the second problem affects all vowels. In general, the transcribers were very unsure about some of their transcriptions.

Initial transcriptions of the imitation of the Saxon /uː/ consisted of open-mid and open (sic!) rounded or unrounded front and central vowels. The perceptual impression had thus been altered to such an extent that it did not seem reasonable to use this imitation for further testing and evaluation. We therefore executed the optimization process again and chose different top-/uː/ candidates, giving more weight to the auditory quality than the degree of formant frequency match. In Table 10.1, only the resulting Saxon /uː/ is presented, omitting the initial version which yielded open-mid and open rounded and unrounded front and central vowels.

The second problem was identified as a problem regarding synthesis mode. During the transcription task it became obvious that the choice of *synthesis mode* (FDS vs. TDS, as discussed in Section 3.1.6, p. 39) had a higher impact on the acoustics of the vowels than expected. We discovered this effect due to perceptual mismatches originating from an unknown source: The automatic playback at the end of an optimization run plays a clear sounding vowel (in the ‘adaptation view’ of VTL, using frequency-domain synthesis), and our first auditory assessment was based on this sound. To generate vowel stimuli of a longer duration, we integrated the corresponding vocal tract target definition in a default gestural score and synthesized the vowel via time-domain synthesis. However, we found that the frequency-domain synthesis mode (FDS) created slightly different perceived vowel qualities than the time-domain synthesis mode (TDS), mainly regarding perceived vowel height. Since vowel height is a central parameter in our articulatory hypotheses of Standard vs. Saxon German pronunciation, this is a very sensitive aspect. Therefore it was a must to ensure that this parameter was not changed by (simple but unconsidered) technical settings of the acoustic rendering.

Separate investigation revealed that one large factor which influenced perceived vowel height was the default setting of the velum. For frequency-domain synthesis (FDS) this parameter is taken from the vowel configuration, whereas in time-domain synthesis (TDS) it is taken from a separate velic aperture tier. Depending on the articulatory height of a vowel, a minimally open velic gesture can introduce nasal qualities that sometimes entail a lower perceived vowel quality (less so for low vowels than for high vowels). While the nasality might be ignored or understood as an ‘idiosyncratic’ articulation, it is hard to ignore the change in perceived vowel height. An analysis of the behavior of F_1 showed that indeed the lowering of the velum had induced an increase in the first formant, which explains the perceptual observations. As a consequence, the audio rendering of all isolated vowels in this chapter was carried out using the frequency-domain synthesis mode to avoid manipulations by external velic parameters. Table 10.1 depicts the transcription of the new (FDS) vowel set.[3]

Formant frequencies

The formant characteristics of the final synthetic vowel set are shown in Figures 10.1b and 10.2b, the corresponding values are provided in Table 10.2 (*‘Best sim.’*). The Saxon vowel imitations are in general more centralized, the rounded vowels are derounded (F_2 increase) and the unrounded vowels are slightly lowered but some do not show large differences between the accents (especially /iː/, as expected). Overall, the acoustic measurements are a good match with the phonetic transcriptions presented in Table 10.1. Additionally,

[3] A more detailed account of the transcriptions can be found in Appendix A.2, including vowel transcriptions across the two synthesis modes, presented in Table A.2. General considerations on the distinction of the two synthesis modes are presented in Appendix A.

Table 10.2: Formant frequencies and error values of the six vowels of the simulation in both accents (*SHG*: Standard High German, *Sax*: Saxon-accented German). Target formants stem from the natural voice recordings of both accents (cf. Section 10.2.1), the corresponding carrier words are written above each vowel's acoustic details. Error values are computed with respect to the target formant frequencies of the natural voice recordings, according to Eq. 4.1 (Section 4.2.4, p. 57). Values of the start configuration (*Start conf.*) represent the acoustics of the corresponding vowel in the default phone set in VTL. They are presented for reference to indicate the acoustic distance of the start configuration to the target before each vowel imitation. Error values of the best vowel simulations (*Best sim.*) indicate the acoustic distance after simulation (printed in bold face). 10 out of 12 vowels show an error value of below 1 % after optimization.

	Standard High German				**Saxon-accented German**			
	F1	F2	F3	Error	F1	F2	F3	Error
		[Hz]		[%]		[Hz]		[%]
/iː/ in *bieten* /ˈbiːtən/								
Target	268	2263	3070		304	2227	3050	
Start. conf.	231	2360	3265	8.98	231	2360	3265	14.72
Best sim.	268	2252	3060	**0.35**	303	2229	3040	**0.24**
/eː/ in *leben* /ˈleːbən/								
Target	320	2262	2797		364	2084	2731	
Start conf.	327	2284	2818	1.61	327	2284	2818	8.18
Best sim.	319	2250	2810	**0.43**	364	2081	2745	**0.32**
/yː/ in *sühnen* /ˈzyːnən/								
Target	291	1732	2163		313	1961	2507	
Start conf.	286	1906	2474	10.18	286	1906	2474	5.15
Best sim.	286	1759	2304	**4.02**	311	1973	2532	**0.76**
/øː/ in *Flöte* /ˈfløːtə/								
Target	340	1499	2232		365	1867	2471	
Start conf.	373	1633	2231	7.12	373	1633	2231	9.24
Best sim.	338	1500	2248	**0.52**	367	1868	2433	**0.97**
/uː/$_{SHG}$ in *Fuhre* /ˈfuːʁə/$_{SHG}$, /uː/$_{Sax}$ in *Spule* /ˈʃpuːlə/$_{Sax}$								
Target	306	686	2282		314	987	2301	
Start conf.	288	854	2283	14.52	288	854	2283	9.04
Best sim.	296	727	2345	**4.24**	313	990	2307	**0.27**
/oː/ in *loben* /ˈloːbən/								
Target	354	715	2487		343	1209	2288	
Start conf.	379	753	2553	5.46	379	753	2553	23.59
Best sim.	355	715	2478	**0.38**	343	1207	2258	**0.75**

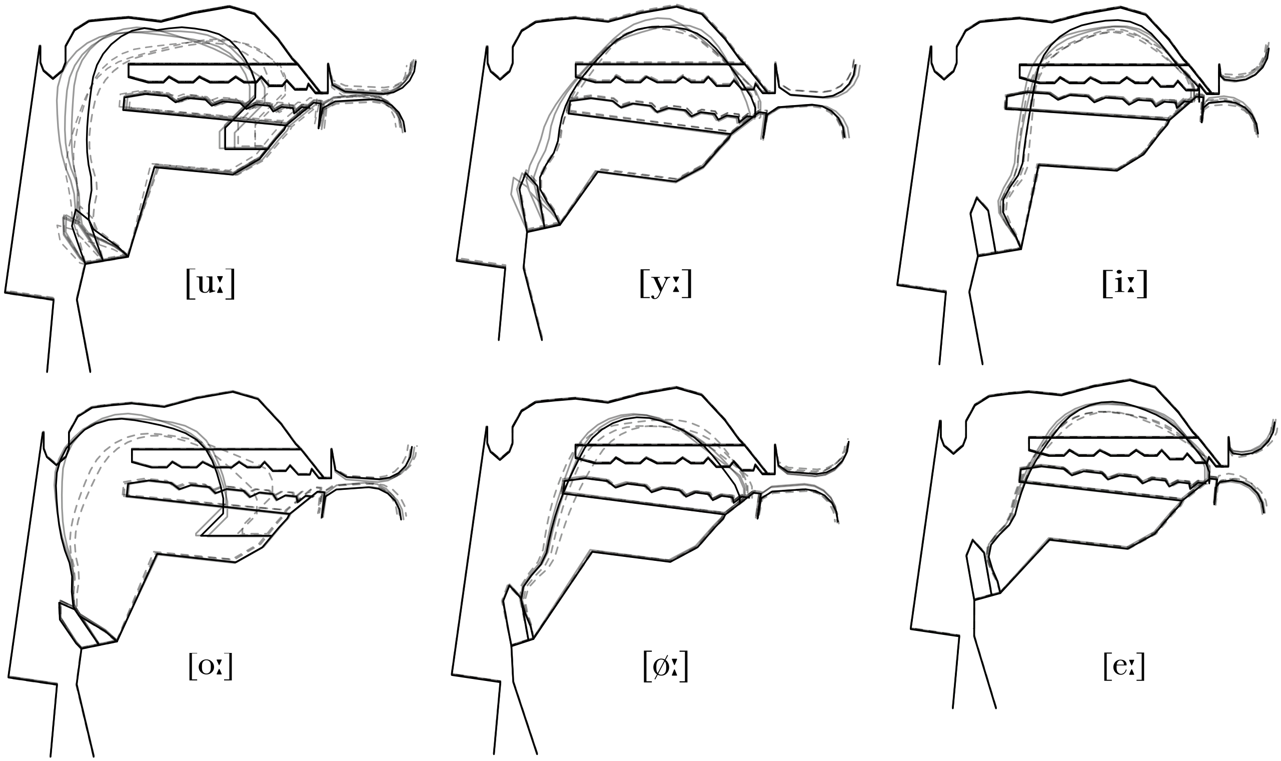


Figure 10.3: Simulated vocal tract shapes for *Standard High German* (solid gray lines) and *Saxon* (dashed gray lines) realizations of /uː oː yː øː iː eː/. Details are explained in Section 10.3.2. For each vowel, the vocal tract configuration in black served as the starting configuration for the optimization function. Details of the optimization procedure are discussed in Section 10.2.2.

the formant values of these imitated vowels seem to be a very good approximation to the original formant values (compare Figure 10.1 (a) vs. (b), and Figure 10.2 (a) vs. (b)). An indication of how much the imitated vocal tract configurations differ from their 'human' target formants can be found in the *Error* columns of Table 10.2, the numbers are based on Equation 4.1 (p. 57).

10.3.2 Articulatory comparison of vocal tract shapes

The multi-layer graphs in Figure 10.3 show that the proposed articulatory settings for Standard vs. Saxon pronunciations are in line with the acoustic descriptions presented above. Solid gray contours outline the simulations of *Standard High German* pronunciation, dashed gray contours show the *Saxon* simulations, while black contours indicate where the start phoneme was located (the one that is stored in VocalTractLab's *standard phone set*, cf. Section 3.1.1, p. 28).

We find that for all vowels under consideration, the articulatory shift takes a similar direction. All Saxon vowels show a fronting and lowering effect when compared to their simulated Standard High German counterparts. We find different degrees of shifting depending on the nature of the target phoneme: The back vowels /uː oː/ show a greater shift than the front vowels /yː øː iː eː/. This confirms the fronting hypothesis because front vowels do not shift backwards (and are constrained from further fronting by the fronted position that defines them). For most rounded vowels, we also find a slight derounding (retraction) of the lips.

The rounded back vowels /uː oː/ show the largest shift from Standard to Saxon pronunciation. The Saxon samples are articulated more anteriorly and, especially the solutions for /oː/, in a lowered manner. The front vowels show articulatory lowering as well, and /øː/ shows some fronting although it is (phonologically) a front vowel to begin with.

Due to the many-to-one mapping problem and the fact that one acoustic output can be generated from different underlying vocal tract shapes, these solutions are only sample solutions accounting for the desired acoustic output. For presentation reasons, only two sample solutions of each accent are depicted in the vocal tract figures. Other solutions, however, show similar systematic shifts indicating that the different solutions from different runs of the algorithm deviate only slightly from one another.

Please note that the cross-sectional area between the lips is set to a minimal area of 20 mm^2 by means of a special control variable (MA$_3$, Table 3.1, p. 29). This is not depicted clearly in the articulatory 2D drawings but it is used in the area function for synthesis.

10.4 Discussion

This experiment concentrated on creating and phonetically comparing two vowel sets from different German accents: the six high and mid-high long vowels /uː oː yː øː iː eː/$_{Sax/SHG}$. Whether the intended accent is robustly perceivable by naive listeners is presented in the following chapter where the vowels are integrated into words. The discussion at this point therefore focuses on the articulatory and technical aspects of the regional vowel imitation, while the discussion in the following chapter brings up aspects such as accent perception and word synthesis.

10.4.1 Phonetic aspects

Regarding the phonetic aspects of this experiment, we found congruent results in both the auditory-acoustic and articulatory analyses, so the acoustic and the articulatory domain can be used together to address the hypotheses of accent pronunciation. As expected, we found general systematic differences between the two accents of German for the vowels under consideration. The main result is that we can confirm prevalent impressionistic and literature accounts of articulatory descriptions of Saxon: Our data suggest the same direction of shift, fronting and lowering, especially for rounded vowels and back vowels. The results may indicate a fronted and lowered *articulatory setting* of the vocal tract for Saxon. To find out whether the produced vowels are indeed identified with their intended accents, the next chapter presents a formal auditory evaluation.

10.4.2 Synthesis technical aspects

Concerning technical aspects, this experiment illustrates that constrained acoustic-to-articulatory inversion carried out by a formant optimization function can be a useful tool to acquire articulatory data for the entire vocal tract in a fast manner. This can be used e.g. to extend the phone set of a model speaker and to compare pronunciation variations.

The optimization method enables us to find, via the acoustic target values, articulatory configurations which are virtually impossible to achieve by manual adjustment. Pretesting showed that manual adjustments following specific articulatory hypotheses, such as fronting and lowering, were impossible to execute in such a precise way that a vowel quality emerged as desired. The articulatory shifts that need to be applied in such a task are too minute for manual adaptation, and interaction with the graphical user interface (GUI) tends to result in over-adjustments, generating strange sounds. Thus, this optimization function is a complementary way of finding plausible articulations for similar sounds compared to the ones already part of the VTL phone set.

Reasons to believe that the data are valid are that, firstly, the different simulation solutions proposed for one particular vowel are in general very similar to one another. This indicates a certain stability of the articulatory system. Secondly, the starting point of the optimization process is located in the close vicinity of the target sound. This particular starting configuration is stored in the default profile of VocalTractLab, which is based on MRI data of a real speaker, and the obtained simulations show similar vocal tract configurations. This suggests that the overall (restricted) domain of the articulatory solutions is based on a realistic configuration. Lastly, the formant error value also gives a rough hint at the quality of the obtained solutions.

With the above considerations in mind, such an optimization-function can be applied as a complementary low-cost method to far more expensive instrumental techniques such as MRI. Again, it seems important to select a starting configuration close to the desired sound, i.e. a phone similar to the one that is going to be imitated. Since the algorithm mainly searches locally, the articulatory solutions will probably be found in the vicinity of the starting configuration. Arbitrary starting configurations are likely to result in a larger formant error than carefully chosen ones.

Three further considerations regarding the results need to be addressed. They concern the lack of co-dependencies of the model articulators, possible effects of body posture during speaking, and speaker normalization. Finally, the scope of the study is discussed.

As discussed in Section 3.1.1, p. 30, most co-dependencies and inter-connections of articulators are not implemented in the model vocal tract. This suggests that the shape of the tongue is over-emphasized in our imitation. Likewise the adjacent articulators (lips, lower jaw) are underestimated, and in a human speaker they would show larger movements than they do in the simulation. This is due to the geometric nature of the vocal tract model in which movements of the tongue have virtually no effect on the lower jaw, and vice versa. So if the optimization algorithm aims to find good solutions, jaw movement does not show sufficient formant frequency changes and is therefore rarely prevalent in the final solutions. On the other hand, moving the tongue immediately produces large changes in resonance frequencies. Therefore the tongue benefits from the setup of the optimization process. As a consequence, we argue that e.g. from an overall lowered position of the tongue one can infer that the jaw also has been lowered, since they are closely inter-connected in human articulation.

Currently, a more refined setup addressing the acoustic-to-articulatory inversion problem is being developed using VTL (Prom-on et al., 2013). It works on continuous acoustic data and implements *embodiment constraints*, “co-adjusting the articulators located near the articulator under adjustment” (p. 350). Thus changes in one parameter also affect surrounding parameters in a physiologically sensible way.

The second consideration concerns the recording setup of the original MRI data for VTL, since studies have shown that body posture affects articulation details (cf. e.g. Steiner, 2010). The data were recorded in a supine position of the speaker. In a supine position, gravity pulls from an unusual angle, which is compensated for by the speaker. Although e.g. Tiede et al. (2000) report that only ‘uncritical’ articulators or parts of them are affected by this gravity effect, the findings nevertheless point out that there are indeed differences in articulation. The acoustic outcome, however, does not differ significantly, therefore it is argued that those vocal tract shapes are as valid as upright ones, simply representing one of the many possible solutions to the one-to-many articulatory-to-acoustic mapping.

In a careful analysis of our experimental setup, this raises the question whether the standard phone set of VTL is ‘colored’ by supine influences, possibly resulting in more strongly retracted articulator positions than those which an upright speaking posture would have yielded. If it were, this could represent a 'slanted' starting point for the optimization algorithm, featuring more retracted articulators (i.e. pulled down when supine) than expected in speech with an upright posture. These considerations are especially sensible in an experimental environment such as the one reported here because we inherently deal with very fine articulatory details and, moreover, our hypothesis about the articulatory shifting in Saxon particularly concerns the horizontal position of the tongue.

However, we believe that the posture effect can be mostly neglected in the interpretation of our results since both accent varieties were simulated in an identical manner and the standard phone set of VTL only provided an articulatory starting configuration in the vicinity of the searched targets and did not further participate in the articulatory optimization of the vowels. This view is supported by a study reported in Engwall (2006) where it is concluded that gravitation can be considered as less problematic than the side effects of sustenation, therefore suggesting to complement static MRI with data from non-static speech productions. This, however, is not something we can address in this experiment since it is more an issue of the core setup of the VTL standard configuration.

As the third consideration, we revisit the issue of normalization. Although we alleviated this problem in the recordings by choosing one 'bilingual' speaker for the natural speech material, there might nevertheless be an issue here. Our speaker produces speech within his vocal tract, and we then imitate his acoustics within the artificial vocal tract of VTL which was adapted to a different real speaker. Implementing our speaker's anatomy into VTL was beyond the scope of this experiment because no anatomical data were available. However, we tried to reduce the dimension of potential physiological mismatches by choosing a similar speaker to the one modeled in VTL (adult and male, and not female or child).

Regarding the scope of this study, we focused on high and mid-high vowels and did so by only considering their static, supraglottal articulatory features. It would be desirable to extend this line of study onto the other vowels of the German sound system to further support our results and to see whether all vowels undergo similar, if not overall systematic shifts in *articulatory setting*. Similarly, other dimensions of AS in Saxon should be analyzed, such as voice quality or prosody. Additionally, since we did not analyze diphthongs or consonantal sounds, temporal properties such as transition times or voice onset times (VOT) have not been considered. Since temporal properties are related to gestural alignment over time, articulatory synthesis seems very suitable to simulate those: The sounds should be straightforward to implement and potential gestural differences should be straightforward to interpret. Adding these dimensions to the stationary aspects of *articulatory setting* as explored in this experiment would, however, complicate the analysis of AS in a multi-dimensional manner since both position and temporal alignment of the different articulators need to be analyzed in a coordinated manner.

In the following chapter, we study how well the imitated vowels are recognized as their intended accent by naive listeners. This complements the current experiment. Furthermore, it introduces technical challenges related to word synthesis in VTL.

Chapter 11

Experiment VII – Words with a Saxon accent

In this chapter, we present a formal auditory evaluation of the synthesized vowels that were the subject of the previous chapter. The phonetic research question is whether those vowels are identified by listeners according to their intended accent: Standard High German (*SHG*) or Saxon-accented German (*Sax*). To make the task more 'natural', we do not present the vowels in isolation and ask naive listeners about their accent rating. We rather embed the vowels in two-syllable carrier words.

Although the phonetic research question is concerned with the perception of accent, we devote a considerable amount of space in this chapter to the creation of word stimuli with VTL. This is the first time in the course of this thesis that numerous words, and not sub-word units or only single words, are generated and used for a listening test. Therefore the process of creating the gestural scores is described, specifically pointing to diverse technical challenges and how they can be addressed to obtain an acceptable pronunciation of the words. The proposed solutions may also be transferred to other words.

The technical research question of this experiment concerns the feasibility of creating a larger set of entire words, and to document the challenges of doing so with VocalTractLab. We expect from previous work that, due to coarticulation effects, creating two-syllable words will require much more effort and will place demands on gestural score design that have not been addressed in previous chapters.

In the following, we discuss strategies for word synthesis in VTL (Section 11.1), explain the word stimulus creation and evaluation procedures of the accent perception test (Section 11.2), and present the results (Section 11.3), followed by a discussion.

11.1 Word synthesis using VTL

In Section 11.1.1 we describe the basic process of creating gestural scores for words. Since they typically need further improvement, a detailed discussion on overcoming synthesis artifacts is presented in Section 11.1.2. Section 11.1.3 sums up the most useful strategies that we found.

11.1.1 Creation of basic gestural scores for words

The creation of a word starts by writing a *song file* which lists all phones of the target word in a predefined format. Pitch and duration values can be added on the syllable level. This represents a rudimentary kind of TTS synthesis and avoids having to create a gestural score from scratch. It works on a syllable basis and was originally implemented to create gestural scores for songs (see also Section 4.2.2, and Birkholz, 2007a). Governed by rules, the necessary gestures are inserted into the tiers of the gestural score. The score still needs extensive adjustment before an acceptable pronunciation is obtained. Nevertheless, using the song file means a considerable speed-up when creating gestural scores for words.

To obtain a first acceptable version of a synthetic word, the word is recorded from a natural speaker and the recording is analyzed with respect to phone durations and pitch contours using Praat (Boersma, 2001). The goal is to copy these rhythmic and intonational patterns. However, transferring these acoustic ‘surface’ values to the gestural score is no trivial task because it is not a direct transfer process. Two main reasons are given here.

Firstly, the phone duration is measured on the signal surface of the word, it thus only provides a rough estimate of the duration of the underlying gestures involved. The duration of a gesture does not necessarily match the duration of its acoustic signal, because depending on the velocity of the articulators or the gestural context, certain acoustic cues surface earlier, others later. For example, vowel gestures can be acoustically obscured by consonantal gestures, thus the vowel gesture can be much longer than the actual vowel sound. This is in accordance with the basic principles of articulatory phonology, where the concept of overlapping gestures offers a basic explanation for the variable acoustic durations of a given phoneme, cf. also Section 3.1.5.

Secondly, to obtain the desired pitch contour, start, end and slope values of the f_0 gestures need to be determined appropriately. The current framework of VTL is somewhat restricting, so it is sometimes necessary to include numerous very short f_0 commands to obtain a sudden peak, or drop, of fundamental frequency.

These first adjustments can be time-consuming and need practice before an acceptable output is generated. It cannot be stressed enough, however, that the adjustments of segmental duration and pitch contour are essential prerequisites to obtain naturalistic sounding synthesis output. These first approximations of rhythm and intonation present a noticeable enhancement of quality regarding naturalness and intelligibility compared to the initial song-file-generated score.

After we have executed the above steps to create basic word stimuli, these stimuli are evaluated by several listeners in a pre-test. The evaluation reveals that the current stimuli need more detailed adjustments to increase the quality of the output in terms of intelligibility and naturalness. The main problems are acoustic artifacts during phone transitions and unexpectedly marked sound qualities, such as the alveolar lateral approximant /l/ sounding too dark (velarized) in back-vowel contexts due to unwanted coarticulation. The artifacts and segmental imprecisions sometimes degrade naturalness and intelligibility to such a degree that the word becomes unintelligible.

To improve the words, the gestural scores are manually adjusted in more detail. This alleviates most of the pronunciation problems. These adjustments are sketched in the following section because they represent typical challenges during the work on word synthesis with VTL, and they may extend to articulatory synthesis in general.

11.1.2 Articulatory adjustments at acoustic artifacts

Up to this point, the gestural scores of the words are based on the output of the song-format gesture function and have been manually adjusted regarding pitch contours and segment durations. The phone targets are taken from the vocal tract configurations stored in the extended phone set of VocalTractLab which includes the different vowel pronunciations created in the previous chapter. During the subsequent adjustment process, durations and pitch contours still receive some optimization and correction, but the core factors for quality boost are located in the following five parameters: Dominance values, articulatory effort, degree to which a target is reached, pulmonic pressure, and phone substitution. Dominance values and articulatory effort are the primary, most globally applicable manipulators, i.e. they enable improvements in many different sound contexts. The others can help primarily in more specific sound contexts.

In a sense, the parameter adjustments really only become necessary due to context effects since the sounds by themselves, in isolation, sound perfectly acceptable. Most of the phenomena are transition artifacts between consonants and vowels. Some phenomena surface acoustically only within one sound but are actually induced by proliferating features from neighboring sounds, and can therefore also be regarded as context effects. In the following, we characterize each of the five parameters listed above and illustrate their effects on samples from our speech material.

Dominance values

Dominance values are specified separately for each articulatory parameter in each phone (see Section 3.1.2), and they stay constant for every use of the target phone. By determining the importance, i.e. dominance of each articulator, or rather each vocal tract parameter, the dominance values shape the coarticulation details between sounds. If the dominance values are not optimized, i.e. made suitable for all contexts, the phones may sound alright in

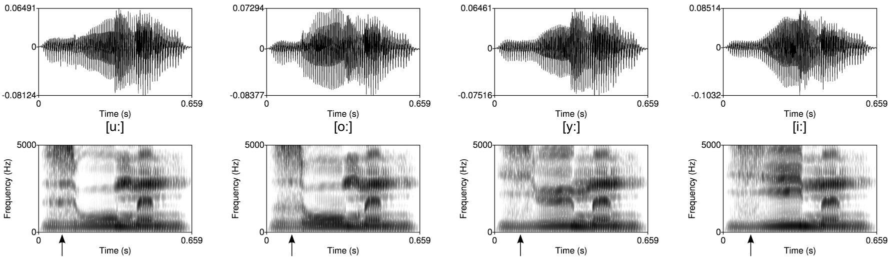


Figure 11.1: Different vowel contexts after /z/ in /ˈz_mən/ (from left to right: /uː oː yː iː/$_{SHG}$). When /yː/ or /iː/ are inserted, the word-initial fricative shows some sort of unwanted ‘lisp’ and palatalization, whereas an insertion of /uː/ or /oː/ produces no artifacts.

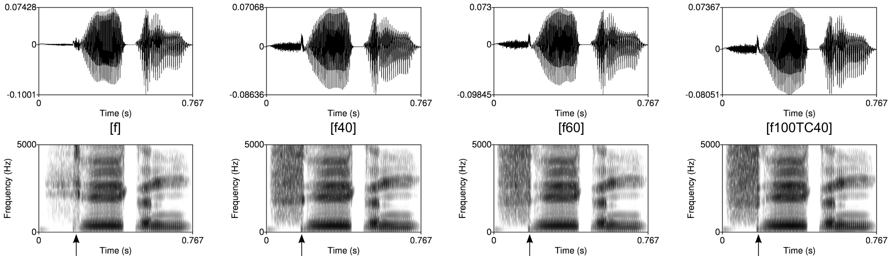


Figure 11.2: Different dominance settings for /f/ in /ˈfyːgən/. From left to right, with increasing dominance: Default /f/ in standard phone set, *f40*, *f60*, *f100TC40*. The default /f/ shows a lot of additional and unwanted friction noise between the first two segments (/fyː/). *f100TC40* reduces this to an acceptably small transitional friction noise which does not distort perception. *f40* and *f60* induce interdental sound perceptions. The complete phone definitions can be found in Appendix F, Table F.5.

some contexts, while in other contexts the sub-optimal dominance settings become audible. This context-dependency is illustrated in Figure 11.1 for the word <zoomen> (‘to zoom’), showing the effect on the initial /z/ of replacing the /uː/ with three other vowels. Please note that an (underlying) vowel gesture always starts at the beginning of the syllable (cf. also Figure 3.5, p. 36). In the carrier word /ˈzuːmən/, the vowel gesture competes with the consonantal gesture for /z/ at the onset of the word. Thus, when calculating the final trajectories, this leads to four different simulated vocal tract configurations during the fricative in the four different vowel environments /uː oː yː iː/$_{SHG}$.

Sample situations, in which an adjustment of dominance values helped, are depicted in Figures 11.2 and 11.3. Their corresponding full parameter listing can be found in Appendix F, Table F.5. The examples involve selective adjustments of a number of articulatory parameters. It often proved effective to, firstly, adjust the dominance values of the tongue center (TC), and secondly, to slightly increase dominance in parameters with originally less

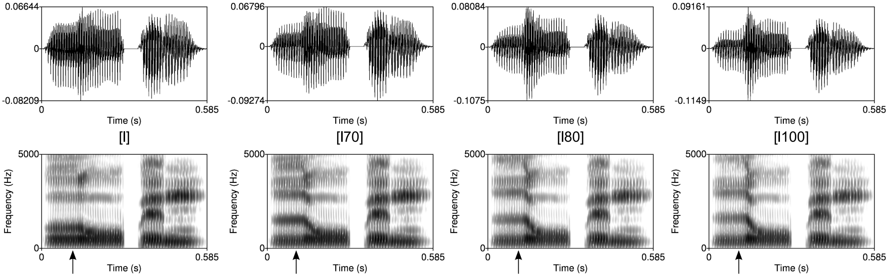


Figure 11.3: Different dominance settings for /l/ in /ˈloːbən/. From left to right, with increasing dominance: Default /l/ in standard phone set, *l70*, *l80*, *l100*. The default /l/ produces a strongly velarized, 'dark' /l/ in this vowel context. With *l100* we obtain a hyper-articulated /l/, which sounds inappropriate and additionally introduces unwanted friction noise. Both *l70* and *l80* sound acceptable. /loːbən/$_{Sax}$ does not make the velarization problem obvious, presumably because the adjacent /oː/$_{Sax}$ is more anterior compared to the standard and *SHG*-accented /oː/. The complete phone definitions can be found in Appendix F, Table F.5.

than 100 %. We repeatedly noted that front vowels and back vowels induced critically different coarticulation patterns. Sometimes, front vowels caused unexpected problems (Figure 11.1), in other situations artifacts were induced by back vowels (Figure 11.3).

Articulatory effort

Articulatory effort is the second main lever for controlling coarticulation in VTL. The values have to be adjusted at every occurrence of the phone on the gestural tier. Articulatory effort values influence the *whole* sound in this *one* instantiation, whereas the dominance values change the influence of individual parameters *within* a sound but for *every* occurrence of that sound. Since articulatory effort defines the velocity with which an articulator moves towards a target, changes in velocity may make it necessary to subsequently adjust the duration of the gesture, or the duration of a neighboring gesture. If a target is reached more quickly, a gesture of the same length produces a longer acoustic segment than when the target is reached more slowly.

Figures 11.4 and 11.5 illustrate how different articulatory effort values (given in Hz) result in different segmental durations and burst intensities. Additionally, manipulation of noise intensity cannot only be done by manipulating the supraglottal, segmental gesture but also the articulatory effort of the glottal abduction gesture. In the word /ˈfyːgən/$_{SHG}$ for instance, we obtain an appropriate /f/ sound but the transition to /yː/ shows an additional friction artifact. We are able to overcome this by maximizing the articulatory effort of the glottal gesture during /f/ from 8 Hz (default) to 16 Hz. A similar transition artifact in the word /ˈʃiːnən/ is removed by maximizing the articulatory values in both the glottal and supraglottal gestures to 16 Hz.

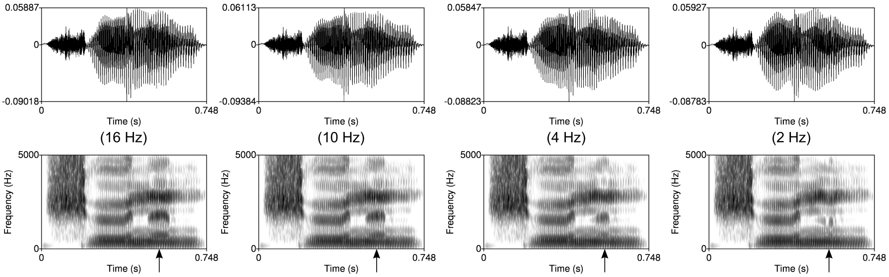


Figure 11.4: Different values of articulatory effort of the Schwa in /ˈʃøːnən/, producing different segment durations. From left to right: 16 Hz, 10 Hz, 4 Hz, 2 Hz. Using 10 Hz sounds best in this context. Higher effort (16 Hz) produces a prolonged Schwa, which is too long and seems too stressed. Lower efforts gradually make the Schwa disappear acoustically.

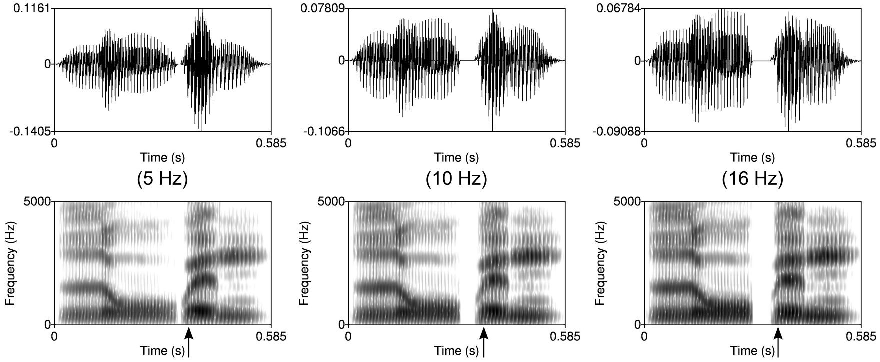


Figure 11.5: Different values of articulatory effort in the supraglottal gesture for the bilabial plosive /b/ in /ˈloːbən/, influencing the precision of the plosive release. From left to right: 5 Hz, 10 Hz, 16 Hz. 16 Hz sounds best, yielding the most precise burst sound. The differences are not large, though, and the best candidate still shows some weaknesses.

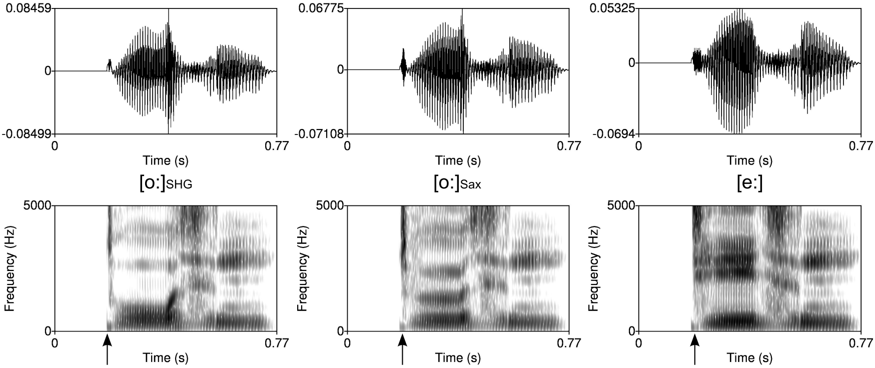


Figure 11.6: Different vowels entail different degrees of burst intensity in the preceding plosive, here /ˈt_zən/ with the vowels /oː/$_{SHG}$, /oː/$_{Sax}$ and VTL default /eː/ (left to right). This can be compensated by varying the degree of glottal opening.

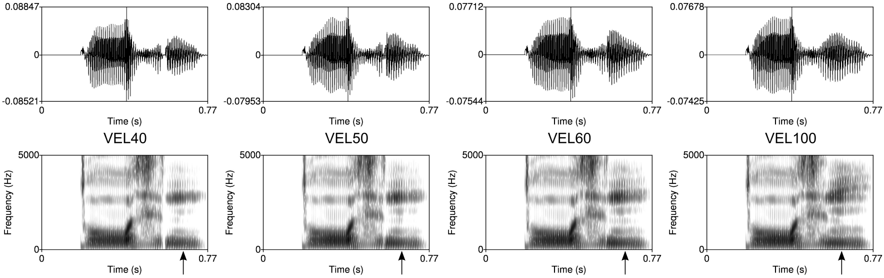


Figure 11.7: From left to right: Increasing degrees of velic aperture for the alveolar nasal in /ˈtoːzn̩/. This entails changes or ambiguities in the perceived place of articulation, in the extreme case (VEL100), a velar nasal is perceivable.

Degree to which a target is reached

The degree to which an articulatory target is reached is defined by the *amplitude* of a gesture, provided there is sufficient time for the articulators to complete the targeted movement. Maximum amplitudes e.g. on the velic and glottal tiers represent a maximum opening at the velic passage and glottis. Figures 11.6 and 11.7 show situations in which glottal and velopharyngeal opening adjustments can be used to optimize burst intensity and nasal quality.

Pulmonic pressure

Pulmonic pressure is the basis for all audible sounds in VTL. For a typical speech utterance it rises at the beginning, then stays relatively constant, and falls at the end of an utterance. To improve particular sounds, pulmonic pressure can be massively increased or decreased for a very short time. There are also interactions with the degree of glottal abduction.

We e.g. increased pulmonic pressure when word-final /n/ was not sufficiently audible otherwise. Pulmonic pressure variations can also be used to adjust plosive bursts and stressed syllables.

Phone substitution

The final strategy to optimize the sound of words consists of experimenting with alternative sounds in the case that artifacts between a consonant and a vowel persist despite the application of the above-mentioned measures. It sometimes helps to replace the first part of the original vowel target – i.e. in the part overlapping the preceding consonant – with a Schwa target. During the overlapping phase, the vowel is not audible until the point of release, it is acoustically hidden and can be replaced, as long as this does not destroy formant loci behavior. Perhaps other vowels work as well, but the Schwa seemed to us the best suited one since it is also the default, or ‘neutral’ vowel gesture in VTL.

This method of sound replacement may be regarded as following the notion of the “trough effect” (Gay, 1975, Perkell, 1986 in Lindblom et al., 2002: 245, see also Fuchs et al., 2004). In VCV-sequences, a trough represents a “discontinuity in anticipatory coarticulation” (Lindblom et al., 2002: 245) during the articulation of the intervocalic consonant. It is supposed to indicate that a neutral position, e.g. of the tongue body, is approximated (Gay, 1975). This challenges the model of coarticulation by Öhman (1967), whose assumptions are present in the implementation of VocalTractLab, claiming that the vowel articulation is continuous and that the consonantal movements are superimposed on it (Lindblom et al., 2002: 249).

The replacement strategy is e.g. applied in the word /ˈfyːɡən/$_{Sax}$ to overcome a strong ‘twittering’ artifact after the fricative. The hidden Schwa has to be of sufficient duration, but not too long as to appear acoustically before the /yː/ and thus producing a diphthong. Combined with adjusted articulatory effort values, the pronunciation becomes acceptable.

11.1.3 Conclusions regarding the tweaking of gestural scores

From a technical-articulatory point of view, we have to deal with pronunciation distortions that stem from different levels of the articulatory control mechanisms of the synthesizer. While some problems are phone-inherent and therefore ubiquitous, most problems with target phones become obvious only in a specific context, such as front vowels vs. back vowels. In summary, we found two main parameters that produce the largest improvements. Firstly, the target's phone definition can be analyzed to identify critical *dominance values*, which are most often too low in the current standard phone set. By adjusting individual dominance values, we can often improve the sound of a phone in all its occurrences on the gestural score. Secondly, distortions can be alleviated by adjusting the default *articulatory effort* values. Suboptimal values lead to inappropriate velocities of an articulator. The articulatory effort values therefore have to be adjusted carefully for every occurrence of a phone. Additional settings such as the amplitude of the gestures are also important. They change the degree to which a target is reached, taking into account that 100% target reaching is not always appropriate. Furthermore, we achieved improvements by adjusting pulmonic pressure and by phone substitution.

Interdependencies between the adjustment parameters make the adjustment task a non-trivial problem. For instance, the duration of an articulatory gesture is not equivalent to the duration of the corresponding acoustic cue. When articulatory effort is increased, a cue may become acoustically too prominent and long, or may change in sound quality, although the underlying gesture was not changed. Similarly, if the articulators move faster, the degree to which a target is reached within a given time may be higher. In such cases, a shorter gesture duration may solve the problem. Thus, adjusting one parameter often makes it necessary to adjust other parameters afterwards.

Although the above-mentioned parameters generally improved our basic utterances to a considerable degree, some distortion phenomena could not be overcome. To illustrate this, we list three of them here. *a)* The burst of the initial plosive in the word /ˈbiːtən/ sounds slurred. Increasing articulatory effort and the degree of glottal abduction produces the impression of a /p/ rather than a precisely articulated /b/. This is in line with the necessary improvements stated in Section 3.3, which include improved burst modeling. *b)* In the last syllable of /ˈtuːtən/ we could not overcome an artifact in the transition between /t/ and /n/, or /t/ and /ə/, if the Schwa was not to be elided. At this point, a more distinct control of nasal burst release would be probably helpful. We decided to keep the Schwa, although it would make the word sound a bit too overarticulated. *c)* The initial fricative in /ˈhuːpən/ was very hard to simulate. It was either too weak, or it induced the perception of being some sort of uvular fricative instead of a glottal one.

The above account of technical articulatory details illustrates the degree of craftsmanship currently involved in creating gestural scores for words. If the arrangement of gestures on the different tiers is done with great care, working knowledge and extensive tweaking, an acceptable output can generally be achieved. However, the great effort needed to get there perhaps indicates a certain 'lack of robustness' currently present in VocalTractLab: Tiny changes of articulatory targets can have a tremendous impact on the acoustic rendering, introducing whistling, twittering noises, unwanted coarticulations, vowel quality changes or vowel elisions, and imprecise articulation in general.

The problem seems to lie in the management of coarticulation so that resulting intermediate (transitionary) states between sounds easily cause distortions. Maybe there are ways to make coarticulation control more robust? A first step of improvement could be to find a way to globally optimize the dominance values for each phone, which seems the most powerful factor in the adjustment issue. The issue is connected, in principle, to the notion of the quantal nature of speech: Minimal changes in one area of the vocal tract may cause maximal overall acoustic changes (Stevens, 1989). So for the sensitive areas in phone transitions, a 'robustness' variable could perhaps be introduced to e.g. prevent articulators from creating too small, friction-critical cross-sectional areas. In parts, this is already implemented for the creation of static sounds where it is e.g. possible to define a minimal cross-sectional area between the lips for vowels ($\mathrm{MA}_{1..3}$, see Section 3.1.1, p. 28).

This situation makes all the more apparent the complexity of relations between the numerous control parameters. In humans, the control takes place on a higher level, with reduced degrees of freedom, and the accompanying adjustments – following the current understanding – are guided by auditory feedback.

These illustrations have highlighted the delicate nature of the mutual, coarticulatory influences of the different parts of the articulatory system. To conclude, we would like to emphasize how important a well functioning coarticulation management seems to be for generating naturalistic speech. Not only is it desirable to achieve realistic and acoustically smooth transitions between segments, but it also seems important to handle coarticulation well in general. The strongest coarticulatory effects are observed in the close vicinity of a segment and are covered by the mechanisms of the current dominance model of VTL. However, it is perhaps also worth incorporating *long domain coarticulations* into the simulations. As has been shown (e.g. Heid and Hawkins, 2000), they extend as far as several syllables from the critical source segment and reliably help listeners in decoding the speech, natural or synthetic, although their acoustic effect is only small.

11.2 Data and analysis

In this section we first describe the goal and hypotheses of the evaluation regarding perceivable accent (Section 11.2.1). Then, the compilation of the stimulus material (Section 11.2.2) used in the perceptual and acoustic evaluation (Section 11.2.3) is described.

11.2.1 Listening test goal and hypotheses

In Chapter 10, the standard phone set of the synthesizer has been extended by the twelve vowels /iː eː uː oː yː øː/$_{SHG/Sax}$. They represent close acoustic-articulatory imitations of one male speaker who speaks Standard High German and Saxon-accented German, using the anatomy of the VTL default speaker. To validate these simulation results, we conduct the following perception test. It aims to find out whether naive listeners recognize the intended accent of the word stimuli, which are described below. The synthetic stimuli are specifically designed to vary only in the target vowel, to evaluate the perceptive power of these vowels. A sound (word) that was intended to function as a Standard High German imitation should be recognized more strongly as 'intended *SHG*' than the corresponding sound that was intended to be imitating the Saxon pronunciation, and vice versa.

We hypothesize that the intended accent is perceived better in some vowels than in others, due to their articulatory properties. The articulatory characteristics of Saxon vs. Standard High German is a fronting and lowering effect. We therefore assume *i)* that the rounded vowels are more distinctly identifyable than the unrounded vowels because the articulatory and auditory differences are larger for rounded than for unrounded vowels. *ii)* Further, we assume that the front rounded vowels are less distinguishable than the back rounded vowels. Since the Saxon articulation shift comprises, where possible, both a lowering and fronting component, the front rounded vowels cannot be articulated more to the front due to lack of forward space. This might be compensated by derounding though (cf. formant characteristics depicted in Figure 10.2). *iii)* Least clear-cut rating results are expected for the unrounded front vowels since their articulatory differences are smallest and the acoustic differences presented in Section 10.3, specifically for /iː/, are also small.

Additionally, the synthetic voice (main condition) is compared to the human voice (baseline condition) in terms of perceptible accent. To evaluate the *synthetic* words, only the vowel is exchanged to mark the accent while the rest of the gestural score is identical for both accents (except for minor adaptations to reduce acoustic artifacts, see Section 11.1.2). This means that other accent-induced influences such as changes in other sounds, especially consonants, or changes in prosody, are not imitated by the synthetic stimuli. In contrast, in the *baseline condition*, the human speaker pronounces the entire word either in Standard High German or Saxon-accented German, thus a lot more acoustic cues are available for a listener to identify an accent. We therefore assume that the accent ratings in the baseline condition (human voice) will be more clear-cut than in the main condition (synthetic voice).

11.2.2 Stimuli

We first describe the generation of the synthetic stimuli (main condition), then the preparation of the human stimuli (baseline condition). While this section focuses on the process of stimulus creation, an overview of the structure of the complete stimulus set is given at the beginning of the next section.

The vowels which are to be evaluated are embedded in carrier words whose selection process is described here. All words are of comparable phonological structure. They are two-syllable German verb infinitives with little consonantal complexity, stressed on the first syllable.

The carrier verbs are taken from the recordings described in the previous chapter (Section 10.2.1), they are listed in Table 11.1. Since whole words are synthesized, the demands on segmental quality are high because the results need to be intelligible. Several candidates proved to be too complicated to synthesize with an acceptable segmental quality and were therefore discarded. The selection criteria are the following: They need to have good audio quality since these words are used for playback in the baseline test condition (human voice). Additionally, they must not show reductions, and they must be intelligible. We test this by asking two naive listeners to write down the words they hear. Intelligible candidate words are then synthesized to find out during several synthesis iterations which ones work best for resynthesis.

Due to constraints regarding segmental complexity, the German verbs are not balanced for occurrence frequency. The list of carrier words complemented with indications of the different frequencies are provided in Appendix Table F.3. To compensate for possible comprehension imbalances, all words are presented in their orthographic form in the listening test before audio playback is started.

The 'basic' word forms in the synthetic voice are created by following the word-synthesis strategies put forward in Section 11.1. The gestural scores obtained in this way contain the default VTL phones and are *not* used in this form in the listening test. Instead, each score is copied twice to become a Saxon and a Standard High German stimulus by replacing the default vowel with the new target vowels /iː eː uː oː yː øː/$_{Sax}$ and /iː eː uː oː yː øː/$_{SHG}$, respectively. As becomes obvious during this procedure, the newly inserted vowels cause different coarticulation effects than the default vowels due to their differing articulation. Therefore, the surrounding gestures and the target vowel gestures themselves are slightly adjusted to reduce the artifacts, again using techniques described in Section 11.1.

The same words that are uttered by the synthetic voice are now also prepared with the human voice. For this, the original recordings are resynthesized. The goal of resynthesis is to degrade the naturalness of the speech signal while at the same time preserving all regional characteristics. Resynthesis is therefore used to decrease the gap in naturalness between the human and the synthetic voice. We use Praat's (Boersma, 2001) overlap-add resynthesis algorithm (PSOLA). To obtain audible degradation of naturalness, time scaling is applied in addition to simple resynthesis. The original recording is time scaled by factor 2, then resynthesized, the output is time scaled by factor 0.5 and again resynthesized, resulting in a resynthesized signal with the original speaking rate.

Table 11.1: Overview of the perception test, consisting of two separate test conditions, each evaluating 36 stimuli in two accents.

Voices (test conditions)	Intended accents	Vowels and carrier words
Synthetic (= main condition) or *human* (= baseline condition)	*SHG* or *Saxon*	/oː/ – loben, tosen, wohnen /uː/ – hupen, tuten, zoomen /øː/ – blöken, lösen, schönen /yː/ – fügen, grüßen, sühnen /eː/ – leben, lehnen, nehmen /iː/ – bieten, schienen, sieden

11.2.3 Analysis methods

The synthetic and human stimuli are evaluated in a formal perception test, which is described first. Then, they are acoustically analyzed.

Perception test design

The evaluation consists of two separate listening tests, the main condition (synthetic voice) and the baseline condition (human voice). An overview of the test design is given in Table 11.1. Participants of the perception tests either take part by listening to the synthetic voice *or* the human voice, not both. In each listening experiment, we use 2 x 6 x 3 = 36 stimuli in total, calculated from 2 intended accents (*SHG*, *Sax*), 6 vowels /iː eː uː oː yː øː/, and 3 carrier words per vowel.

Each listener is presented with only a subset of these stimuli to reduce the test duration for any one subject. In total, the items are distributed in such a way that the occurrence of words, vowels (phonemic quality, roundedness, back/front tongue position) and intended accent is balanced across the whole test and within each participant. The items are randomized with the restriction that no two consecutive items contain the same base vowel.[1] In a series of six items, each base vowel occurs exactly once, with three of the base vowels occurring in one intended accent, three in the other. The participants read the orthographic version of the target word and can listen to the audio up to ten times. Each participant hears all 18 wordings, and then again their first 6 wordings, but each now with the opposite accent. Thus, after a group of 3 participants, all stimuli have been rated exactly twice.

During the first half of the test, a listener has to answer the question “How Saxon does the word sound?”, while in the second half they are asked “How Standard High German does the word sound?” The sequence of these questions is balanced across all participants. The reason for not providing a single rating scale labeled “Saxon” at one end and “Standard German” at the other is that it makes it impossible to unambiguously identify answers that mean “this word sounds neither Saxon nor Standard German.” We do not want to rule out this possibility because by using articulatory speech synthesis, we have no orientation regarding

[1] The term base vowel denotes the vowel phoneme regardless of accent.

its overall performance. The two different questions give the theoretical possibility of a stimulus to receive the lowest rating for Standard *and* Saxon German. This would suggest a third attribute of the stimulus, be it "some other accent", or, more probably, a "strange stimulus quality." The latter would probably interfere with the accent rating proper.

The ratings are entered on a 7-point scale, labeled "Very High German" ("sehr hochdeutsch" = 7) and "Not High German" ("nicht hochdeutsch" = 1) at the ends, or "Very Saxon" ("sehr Sächsisch" = 7) and "Not Saxon" ("nicht Sächsisch" = 1), respectively. Screenshots of the corresponding task prompts can be found in Appendix Section B.5.2).

In addition to asking for an accent rating, we ask for an indication of how confident a participant feels about their answer, also on a 7-point scale, labeled "Very sure" (7) and "Not sure" (1) at the ends.

Both listening tests (main and baseline condition) are performed via the world-wide web by means of a web-based test interface (Draxler, 2011). Screenshots of the phases of the experiment can be found in Appendix Section B.5.2, Figures B.3 to B.6. General remarks on web-based testing are presented in Section 4.2.7. The test was available for approximately four months. During this time, we invited people via email and newsletter postings to participate and distribute the invitation further. The main motivation for conducting this listening test on the web was that it enabled us to easily collect data from different regions of Germany. The goal was to acquire participants with different regional or dialectal backgrounds because the task was accent-related and the language (accent) background of the listeners may influence the ratings.

Statistics

In total, the web-based test for the synthetic stimuli was executed by 120 participants, the one for the human stimuli by 34 participants. The baseline condition was designed only as a control condition, therefore we collected less data for it. Not all data sets are included in the analysis, though, as is explained in the following. The reasons are mother tongue, repeated participation, incomplete data sets, reaction-time outliers, and rating outliers.

Five participants are excluded because they indicated that they were not German native speakers.

By cross-checking the answers on the demographic questionnaires, we identified a number of participants who completed both tests. Due to this repeated participation, we excluded seven data sets, always keeping the first test of each participant. This was checked by an inspection of the time-stamps in the log files.

In ten cases, the participant did not complete the test. The percentage of incomplete data sets with respect to the total number of participants is called the drop-out rate (cf. Section B.3). Our overall drop-out rate amounts to about 6 % (10 of 154). Drop-out reasons can be diverse, in some cases feedback comments indicated technical problems, in others, one can only speculate. It is noteworthy, however, that dropouts were only found in the main condition, perhaps because of the nature of the synthetic stimuli.

Thirty-one participants (24 in the main condition, 7 in the baseline condition) are excluded because their reaction times for at least one stimulus was above the reaction time threshold (RT_{thr}), defined as $log(RT_{thr}) = log(RT_{mean}) + 2SD$, or it was below 200 ms, which seems to indicate random fast clicking on the answers.

Three participants are excluded as rating outliers. We exclude participants whose ratings average to a value of above 6 or below 2, referring to a rating scale from 1 to 7. Those participants did not use the scale as intended. However, it could also be argued that this may be an indication of the stimuli being too similar or the task too difficult. This may be supported by the finding that we found this type of outlier only in the main condition (synthetic voice).

In total, the analyses are based on 98 data sets: 74 in the main condition (*synthetic voice*, 46 female, 28 male, mean age of 30 years, SD = 8 years, range of 47 years, from 19 to 66), and 24 participants in the baseline condition (*human voice*, 16 female, 8 male, mean age of 28 years, SD = 8 years, range of 32 years, from 19 to 51).

From the demographic questionnaire we extract information about the accent background of the listeners, based on language and geographic data they provided, and split up the listeners into two groups of different 'language backgrounds'. One group has or had exposure to Saxon and other East Central German (*ECG*) dialects, such as Thuringian. These dialects are perceived as sounding very similar, so it seems reasonable to pool the listeners into one group. The other group has not had extensive exposure to Saxon or other ECG dialects. In the main condition of our experiment (synthetic voice), 23 participants belong to the ECG group and 51 to the non-ECG group. In the baseline condition (human voice), 16 listeners have an ECG background and 8 do not.

We calculate linear mixed-effect models (REML, α = 5 %) with lexeme (wording) and subject as random factors using the software package JMP (SAS Institute Inc., 1989–2013). Further descriptive statistics are calculated using the software package R (R Core Team, 2013).

Acoustic measurements

Besides the perceptual evaluation of the entire words, we measure the stationary part of each vowel token as it appears in the final stimuli. The formant frequency analysis is done with the same (standard) parameters as described in Section 10.2.1, using Praat.

These individual measurements complement the formant plots presented in the previous chapter (cf. Figures 10.1a, 10.1b). They were based on one target candidate for each vowel (human voice), and the best-candidate output of the imitation algorithm (synthetic voice), respectively. More importantly, the synthetic voice measurements were obtained from the vowels in isolation, whereas now they are obtained from the vowels in context. Therefore, this acoustic analysis aims at capturing potential effects that the phonetic context may have on the acoustic structure of the synthetic vowels.

11.3 Results

We first present the results of the perception test, starting with general considerations, followed by overall accent ratings (perceived accent) and ratings for each vowel, ratings of each group of lexemes which carry the same vowel, and overall confidence ratings. Secondly, we present the acoustic analyses.

11.3.1 General considerations

Two general considerations about the accent ratings are put forward here, firstly regarding effects of rating scale, and secondly regarding effects of test condition (voice).

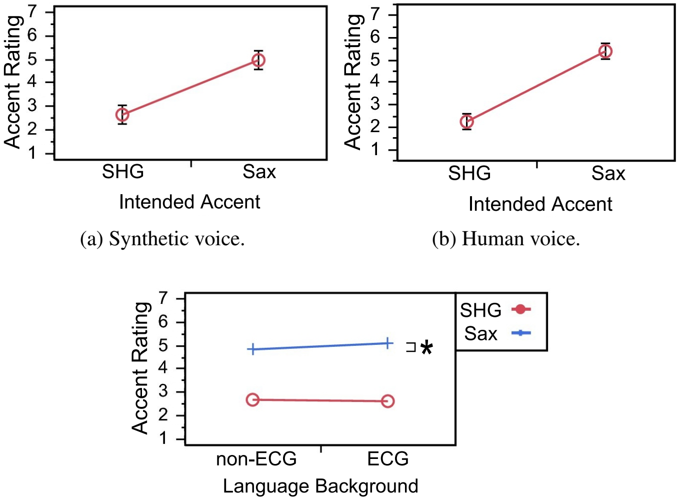
As described in Section 11.2.3, the participants rated half their stimuli using *one* question, and the second half using the other. The meaning of the associated rating scale depends on the question: On a scale from 1 to 7, a rating of 7 for the question type 'High German' means this stimulus sounds very High German to the participants. In contrast, for the question type 'Saxon', a rating of 7 means this stimulus sounds very Saxon. These different versions of the rating scale were used to allow a stimulus to be rated for instance "neither Saxon nor High German."

Analysis of the data however shows no significant influence of rating scale in either of the voices (synthetic: F(1, 1682)=0.0571, p=0.8112, human: F(1, 536)=0.5200, p=0.4712). In other words, the ratings across question types mirror each other. Therefore, the presentation of results uses a 'consolidated' rating scale: We keep the original rating scale for the question type 'Saxon' as is, and the values that were originally filled in as a response to the question type 'High German' are mapped onto this scale by reversing their original scale, i.e. $7_{HighGerman}$ becomes 1_{Saxon}, and $1_{HighGerman}$ becomes 7_{Saxon}. Since the consolidated rating scale incorporates both question types, it can be labeled as follows: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German).

Regarding test condition (voice), we find an overall effect on the accent ratings, which interacts with the intended accent of the stimuli (F(1, 2235)=28.7347, $p<0.0001$). More specifically: For the intended Saxon stimuli, the voice has a significant effect on the accent ratings[2] (t=-2.507, p=0.0133). For intended High German stimuli, no effect is found (t=1.864, p=0.0644).

Across voices, it can be noted that the answers are more extreme, i.e. more distinct for the human voice than for the synthetic voice. This may be due to the strangeness and remaining distorting components of the synthetic voice and the fact that the human-voice stimuli are pronounced completely as *Standard High German* or *Saxon* words, whereas the synthetic stimuli differ only in the intended accent of the target vowel and no other segment. We pick up on this in the discussion (Section 11.4).

[2] t-values $> |2|$ indicate significant results.



(a) Synthetic voice.

(b) Human voice.

(c) Synthetic voice, interaction of accent rating with language background of the listeners (ECG = East Central German).

Figure 11.8: Overall accent ratings (least square means) for both voices, by intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German).

11.3.2 Accent ratings overall and per vowel

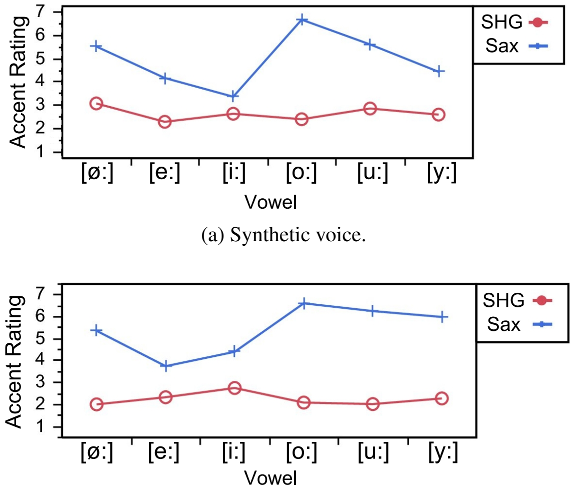
The main hypothesis of the perception test is that intended Saxon stimuli are rated more Saxon than intended High German stimuli, and vice versa. At the same time we assume that the ratings differ depending on the base vowel. Finally, an influence of the listeners' geographic background on the ratings is assumed. We find a significant main effect of intended accent in each of the voices, and a significant interaction of intended accent and vowel. Furthermore, for the synthetic voice only, we find a significant interaction of intended accent and geographic attributes of the listeners. All other interactions are not significant. Details are shown in Table 11.2.

The mean accents ratings across all stimuli are presented in Figure 11.8a-b, the interaction diagram of the synthetic voice is presented in Figure 11.8c. The corresponding values are provided in Appendix Tables F.6 and F.7.

The mean ratings of each of the six target vowels of each voice is shown in Figure 11.9. Corresponding values are provided in Appendix Table F.8. Each intended Saxon vowel is rated significantly more Saxon than its intended High German counterpart. For contrast test details, see Appendix Table F.9.

Table 11.2: Influences of intended accent (intacc), vowel and listeners' language background (ecg) on the accent ratings. Significant main effects and interactions are printed in bold face.

Source	Synthetic voice		Human voice	
intacc	**F(1, 2147)=1175**	**p<.0001***	**F(1, 679)=908**	**p<.0001***
vowel	F(5, 12)=2.35	p=0.1035	F(5, 12)=1.75	p=0.1958
ecg	F(1, 937)=0.34	p=0.5615	F(1, 29)=1.08	p=0.3066
intacc*vowel	**F(5, 2147)=50**	**p<.0001***	**F(5, 679)=26.5**	**p<.0001***
intacc*ecg	**F(1, 2147)=5.81**	**p=0.0160***	F(1, 681)=0.12	p=0.7345
vowel*ecg	F(5, 2148)=0.23	p=0.952	F(5, 679)=1.53	p=0.1771
intacc*vowel*ecg	F(5, 2147)=0.93	p=0.4623	F(5, 681)=0.74	p=0.5913



(a) Synthetic voice.

(b) Human voice.

Figure 11.9: Mean accent ratings (least square means) for each target vowel in both voices, by intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German).

When examined along the articulatory dimensions [±round], [±front] and [±high], the picture is not very clear-cut and only reveals *tendencies* in the ratings (for details and subgroups see Appendix Figures F.1 and F.2). As hypothesized, intended-Saxon *rounded* vowels are more distinctly rated as “Saxon” than the intended-Saxon *un*rounded ones in both the synthetic and the human-voice stimuli. For [±front] and [±high], we cannot see any meaningful patterning tendencies, making a detailed articulatory interpretation of the results impossible.

11.3.3 Accent ratings of each lexeme

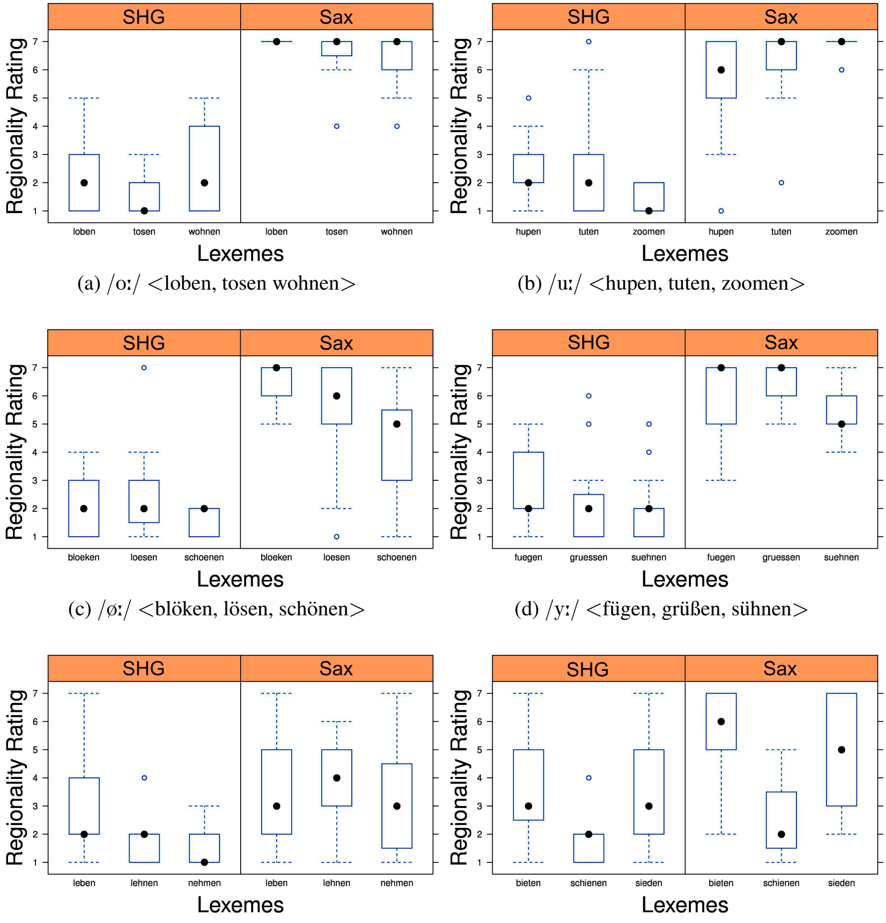
One possible reason why the articulatory dimensions do not show a clear pattern may be the nature of the carrier words. The accent ratings for both voices show an overall significant interaction of intended accent and carrier word (synthetic: F(17, 1677)=18.6954, $p<0.0001$, human: F(17, 526)=9.7442, $p<0.0001$).

The ratings for each carrier word per vowel are shown in Figure 11.10 (human) and 11.11 (synthetic).[3] Statistical details on the behavior of the words within one group are provided in Appendix Tables F.10 and F.11. We first discuss the human-voice results in detail because, as we have assumed previously, they show more clear-cut patterns than those for the synthetic voice. In this sense, the human voice can give us a hint at what was presumably the clearest possible pattern in the ratings. However, both voices show a similar rating tendency.

A rating pattern that comes close to the assumed ‘ideal’ rating can be found for the /oː/-words produced by the human voice, see Figure 11.10a. The intended accent is distinctly recognized in the accent ratings, and all carrier words show a (more or less) homogeneous pattern. In general, the expected ratings according to intended accent can be seen more clearly in the graphs of the carrier words with the *rounded* vowels /oː uː øː yː/, Figure 11.10(a)-(d), than in the graphs with the *unrounded* vowels, Figure 11.10(e)-(f). The carrier words of the vowel /eː/ are well recognized in their *Standard High German* versions, but not so well in their *Saxon* versions. For the /iː/-words the influence of intended accent is even smaller and less homogeneous.

In the synthetic voice, the lexemes with the four *rounded* vowels are again rated most distinctly according to their intended accent (Figure 11.11(a)-(d)). This is again clearest for /oː/. The /eː/$_{Sax}$-words (Figure 11.11e) only show a slight tendency for expected accent rating, and the /iː/-words (Figure 11.11f) receive relatively high Standard German ratings for *both* intended accents.

[3] All box-whisker plots show the mean as a dot and the 25-75 quartile distribution as a box. The whiskers show the minimal and maximal quartile distribution. Potential outliers are depicted as small circles.



(a) /oː/ <loben, tosen wohnen>

(b) /uː/ <hupen, tuten, zoomen>

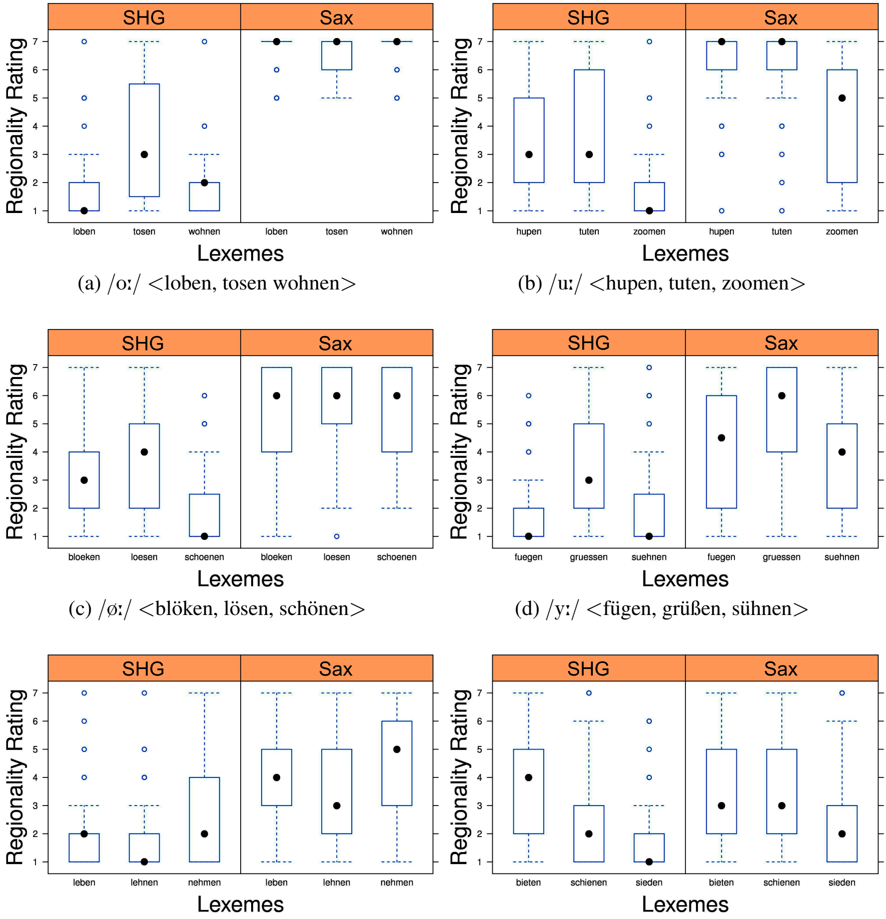
(c) /øː/ <blöken, lösen, schönen>

(d) /yː/ <fügen, grüßen, sühnen>

(e) /eː/ <leben, lehnen, nehmen>

(f) /iː/ <bieten, schienen, sieden>

Figure 11.10: Mean accent rating of each lexeme (repeated twice, from left to right), **human voice**, grouped by target vowels and split by intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German).



(a) /oː/ <loben, tosen wohnen>

(b) /uː/ <hupen, tuten, zoomen>

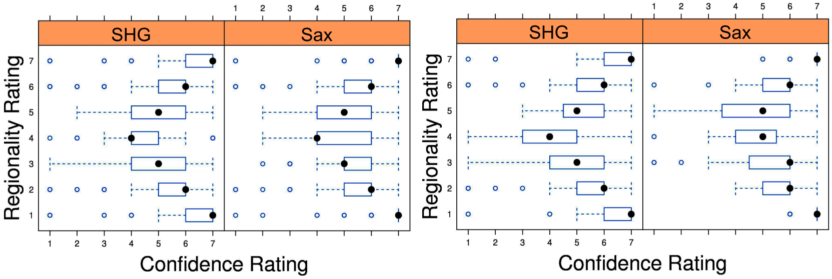
(c) /øː/ <blöken, lösen, schönen>

(d) /yː/ <fügen, grüßen, sühnen>

(e) /eː/ <leben, lehnen, nehmen>

(f) /iː/ <bieten, schienen, sieden>

Figure 11.11: Mean accent rating of each lexeme (repeated twice, from left to right), **synthetic voice**, grouped by target vowels and split by intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German).



(a) Synthetic voice

(b) Human voice

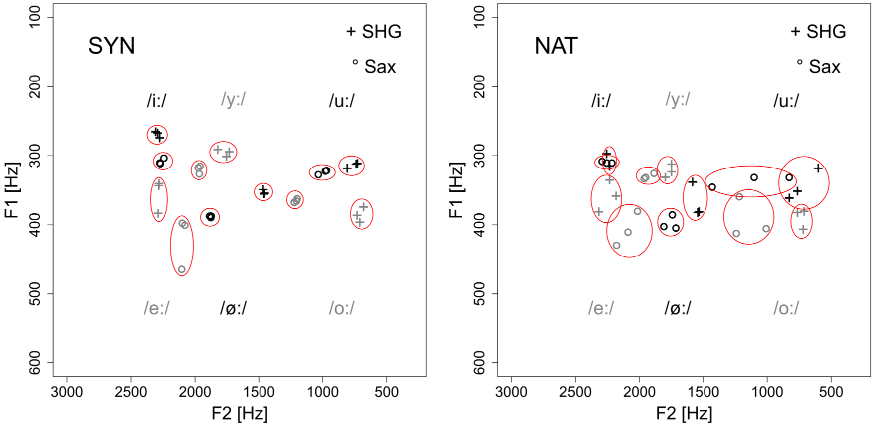
Figure 11.12: Confidence ratings in conjunction with accent ratings, for each voice and intended accent. High confidence values (close to 7) tend to be found in conjunction with distinct accent ratings (close to 1 or 7).

11.3.4 Confidence ratings

For each stimulus, the participants had to give an indication of how confident they felt in providing the accent rating. Figure 11.12 shows the confidence ratings in conjunction with the accent ratings for all stimuli over all participants.

As might be expected, the more distinct, i.e. closer to the ends of the scale, the accent ratings are that a participant gives to a particular stimulus, the more confident the participant is in giving this answer. At the same time, when the accent rating is close to the mid point of the 7-point scale, the confidence is rather low (synthetic: F(1, 1547)=718.59, $p<0.0001$; human: F(1,422)=238.58, $p<0.0001$). This can be seen as an indication that the participants executed the task in a careful way instead of blindly clicking on answers. In the same analysis, we also find an effect of intended accent (synthetic: F(1, 1689)=27.36, $p<0.0001$, human: F(1, 538)=14.29, p=0.0002), but no interaction of the two factors (synthetic: F(1, 1586)=3.70, p=0.0545 , human: F(1, 446)=2.90, p=0.0893). The intended-Saxon stimuli receive slightly higher confidence ratings than the intended-SHG stimuli (least square means; synthesized: Confidence$_{intSax} = 5.84$, Confidence$_{intSHG} = 5.60$; human: Confidence$_{intSax} = 5.92$, Confidence$_{intSHG} = 5.61$).

We report these ratings since we regard the confidence ratings primarily as a control measure for the assumed seriousness with which the participants took part in the listening experiment. From a participant's point of view, we assume that the opportunity to document a possible lack of confidence can contribute to a higher motivation to complete the entire experiment. The participant can make a statement of low confidence rather than feeling frustrated because the task is perhaps relatively difficult for certain stimuli.



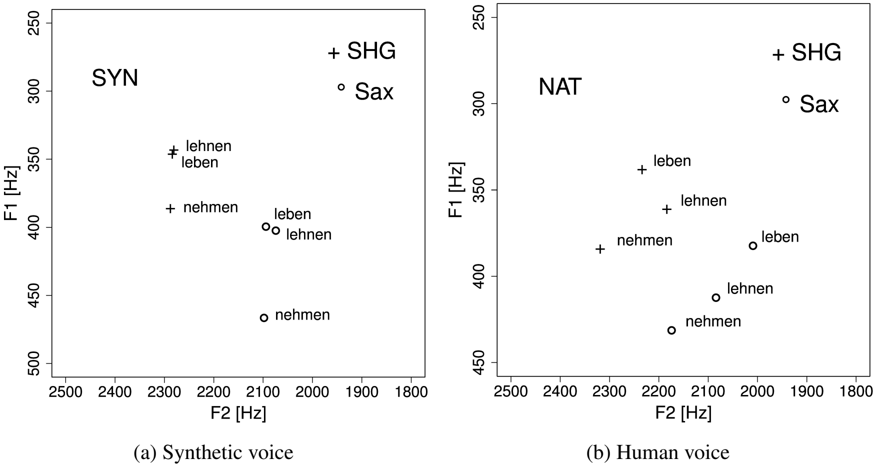
(a) Synthetic voice (b) Human voice

Figure 11.13: Formant frequency plots, F_1 vs. F_2, of the vowel tokens in the final stimuli, by voice and intended accent (SHG = Standard High German, Sax = Saxon). Tokens of the same target vowel and the same intended accent are manually surrounded by an ellipsoid for better visibility.

11.3.5 Formant frequency values

The F_1-F_2 plots in Figure 11.13 show the formant values during the stationary parts of the target vowels in each stimulus in both test conditions. While the general picture shows a relatively good match between the voices, two aspects of divergence are briefly pointed out. Firstly, the human-voice vowel tokens show a greater within-vowel variability than the synthetic ones. The greater variability in the human-voice vowels can be expected due to natural variability and different phonetic contexts. In contrast, the smaller distributions of the synthetic tokens can be expected since all stimuli of one target vowel were synthesized using the same underlying articulatory target definition. Secondly, the acoustic areas of the synthetic vowel targets are more distinct from each other, while the human-voice tokens show some overlap. This is because we selected relatively distinct human-voice vowel tokens to serve as the targets of the constrained acoustic-to-articulatory inversion.

One particularity in the results can be seen in the relatively broad distribution of the synthetic words carrying $/eː/_{SHG}$ and $/eː/_{Sax}$, which is also depicted in Figure 11.14 and seems to be connected to nasalization processes. In the synthetic stimuli, highest F_1 values are found in <nehmen>, while <leben> and <lehnen> show low F_1 values in both imitated accents. The human-voice /eː/ tokens also show highest F_1 values in <nehmen>, <lehnen> shows mid-high values of F_1 and <leben> the lowest ones. These acoustic



(a) Synthetic voice

(b) Human voice

Figure 11.14: Detail formant frequency plots of the /eː/ vowel tokens in the final stimuli, by voice and intended accent (SHG = Standard High German, Sax = Saxon).

observations can probably be explained by the immediate nasal context, which induces a certain degree of nasalization in the target vowel. It has been shown that vowel nasalization has acoustic effects in the region of F_1, which is associated with vowel height perception. Depending on linguistic experience, listeners may compensate for this effect, see e.g. Goodin-Mayeda (2011). cf. also Hawkins and Stevens (1985) for an overview.

Inspection of all supraglottal spatial movements of the virtual speaking apparatus during the pronunciation of the /eː/ targets indeed shows that the only divergence is located in the movement of the velum. It is induced by the nasal context both before and after the vowel and causes the artificial velum to stay below a critical threshold during that vowel. The position of the velum has a considerable influence on the vowel quality of these close-mid vowels. When the velum is lowered (as in <nehmen>), it acoustically changes the target vowel almost into an open-mid vowel [ɛ]. This can be due to the altered oral tract area function since the velum is taking up some space in the oral tract, but it is more likely due, primarily, to the added nasal coupling. This acoustic behavior may indicate, in a positive manner, how sensitive the synthesizer is with regard to imitating natural articulation.

In the gestural scores, the threshold for nasalization is passed when sufficiently large velic gestures are present both before and after the vowel, as in <nehmen>. If they are only present after the vowel, as in <lehnen>, the nasalization effect does not take place and the formants of the /eː/ are similar to those in completely non-nasal surroundings such as in <leben>. This differs from the natural recordings where a one-sided nasal context induces mid-high F_1 values.

This nasality-related, acoustic-articulatory finding is partly reflected in the previously mentioned accent ratings of the synthetic words with /eː/, which show inhomogeneous rating patterns that might be due to the different nasalization contexts in the individual words (cf. Figure 11.11e). Regarding the accent ratings of the human-voice stimuli with /eː/ (cf. Figure 11.10e) it seems that the contributions of the vowel formants to perception are less prominent. We assume that this is the case because the words as a whole are spoken either with a *Standard* or *Saxon* accent, so other influences such as consonant articulation or intonation can come into play.

11.4 Discussion

We ran a web-based listening test to find out whether carrier words are perceived according to the intended accent of the embedded target vowels /uː oː yː øː iː eː/$_{SHG/Sax}$. The successful perceptual assessment validates the target configurations of the vowels that have been described in the previous chapter. We first discuss phonetic aspects of this experiment, followed by technical considerations.

11.4.1 Phonetic aspects

This experiment on words pronounced with different accents puts special emphasis on the 'fine articulatory detail' of word pronunciation and shows how even minor articulatory variations can induce e.g. a changed accent perception.

The overall results indicate that the vowels, integrated in two-syllabic infinitive verb forms, are successfully recognized as "Standard High German" or "Saxon" according to their intended accent. The main deviations from the mean ratings are probably due to the effects of single stimuli, e.g. word forms that were difficult to synthesize with the articulatory synthesizer or words that are infrequent in German. The latter had been part of the test because of their simple segmental structure.

Based on the impressionist articulatory accounts and the results of articulation simulation in the previous chapter, Saxon may be pronounced using fronting and lowering in the speech tract. We therefore expected relatively distinct ratings for rounded vowels and even more distinct ones for *back* rounded vowels since they show the largest differences to Standard German. /oː/ indeed showed the most clear-cut ratings, with /uː øː yː/ being also relatively distinct. Since in unrounded front vowels we do not have much room for fronting, this dimension is not available to make Saxon /iː/ and /eː/ much different from the Standard pronunciation. The slight lowering that took place was obviously only a rather weak cue, so the results for the unrounded front vowels were less clear-cut. Moreover, in some words a (perceptually) lowered vowel was introduced due to a nasal segmental context. Since this led to a stronger Saxon rating even in the *Standard German* stimuli, we assume that the lowering component is indeed a sign for Saxon pronunciation. However, this rating might also mean "sounds strange, is probably Saxon."

As implied in the last sentence, it can be argued that the ratings on the "Saxon" rating scale could be interpreted as ratings 'against proper Standard German' pronunciation, i.e. when a word sounded 'strange' (regardless in which way) it would receive higher "Saxon" ratings. However, the "Saxon" ratings are significantly higher for Saxon words when stemming from listeners *with* exposure to East Central German (ECG) dialects than those of listeners with a *non*-ECG background. This implies that the listeners with ECG exposure seem to identify their familiar accent. The fuzziness of the question design can be counteracted by posing slightly different questions with future test material, such as giving three accents as a choice of answer. Since this was the first time this kind of vowel material was created and tested, we decided to design a rather restricted test. However, as explained above, by using two questions throughout the listening task, we at least provided an opportunity to rate words as "non-Saxon" *and* "non-Standard" – as one single scale would have been ambiguous in this respect.

The ratings may also have been influenced by playback problems that occurred irregularly with some participants and were reported in the final feedback text box. The words sometimes showed short crackling noises towards the end of the playback. While this obviously irritates a listener and may influence the rating, it usually took place a few phones after the target vowel was heard. Therefore, it presumably did not disrupt the target vowel but may still have influenced the overall impression of the word.

Finally, two considerations about the word material of the two tests should be noted here, the first regarding results within the human-voice word set, the second regarding results between the human vs. synthetic word sets. Firstly, some of the less clear-cut ratings between Saxon and Standard within the human-voice word set could be explained by the fact that our speaker might have shown a slight Saxon coloring in his Standard German pronunciation. To check this, independent listeners gave feedback on some words after analysis of the test results. They indicated that e.g. the *SHG* variant of <loben> had a slight touch of a Saxon accent in it. Therefore, some words received Saxon ratings already on the Standard variant.

The second consideration is related to the hypothesis that the human-voice stimuli would receive more clear-cut ratings than the synthetic ones. This hypothesis was primarily motivated by the technical fact that the artificial words contain *only* one single vowel as accent cue, whereas the human-voice words are complete Saxon or Standard pronunciations with no internal manipulations. A second factor that could have contributed to these clear-cut ratings for the human voice is that a diphthongization tendency occurs in some words. This concerns mainly words with the target vowels /eː/ and /oː/ where the speaker shows a tendency to pronounce [eː͜ɪ] and [oː͜ʊ]. Since the artificial vowels are strictly kept as monophthongs, the diphthongization cue is not available in the synthetic stimulus set. The latter issue in itself indicates how important it is to carefully select the speaker for the recording.

11.4.2 Synthesis technical aspects

Regarding the technical aspects, this experiment showed that not only single-vowel stimuli or individual paralinguistic vocalizations, such as laughter, but also ‘regular’ linguistic units such as words can be systematically varied in articulation with VTL. We illustrated how we can generate tailor-made word stimuli to study articulatory effects on perception.

In contrast to the synthesis of laughter, the synthesis of actual words of a language faces much stricter segmental needs, or expectations on the part of the listener. Therefore, it more easily becomes obvious when some segments are not well articulated. In the pilot tests we found e.g. that the intelligibility was too low when no orthographic representation was provided. This motivated the effort we took in the articulatory ‘tweaking’ of the gestural scores, to ensure a basic pronunciation quality that met the expectations of naive listeners. To be on the safe side, we still provided the words in written form.

It became clear that the ‘robustness’ of VTL regarding the co-articulatory effects in words is still relatively low. The acoustic output is susceptible to minimal misconfigurations e.g. in gestural alignment, vocal effort and dominance values, as has already been discussed in Section 11.1.3. It was difficult to obtain the fine articulatory details that were needed for ‘regular’ word pronunciation. However, we found techniques to alleviate the majority of the problems we encountered. Effective measures included adjustments in dominance values and vocal effort. Some articulatory challenges, though, could not be solved. These included e.g. the quality of plosive burst releases, the quality of the glottal fricative /h/, and the possibility for more explicit control over nasal releases. Since the segmental challenges are subject to ongoing research it can be expected that they will be handled more robustly in future versions of VTL (Birkholz, 2013a).

It would be desirable to extend this line of articulatory accent investigation to more sounds of the German sound inventory. The connection of articulation with acoustics in VTL proved to be a valuable method for articulation research, using restricted-scenario acoustic-to-articulatory inversion combined with a visual analysis of the suggested articulatory solutions. In this respect, VTL can be used as a complementary method to the instrumental techniques that record actual articulation patterns (cf. Section 2.1.3).

The results can furthermore build a basis for accented-speech synthesis. Given the large amount of manual work though, the current manual method is more suited to explore single articulation phenomena related to accent rather than building entire new voices. Accordingly, other synthesis techniques are currently more advanced in this respect, based on automatic evaluation of large acoustic corpora. Regarding German varieties, there have e.g. been projects on building acoustic imitations of the Viennese sociolect using HMM synthesis (e.g. Pucher et al., 2010).[4] These applications are, in principle, also possible in articulatory speech synthesis. To make reasonable progress, however, it would be practical to, at least in parts, apply e.g. automatic learning techniques, provided that suitable articulatory corpora are available.

[4] Multi-dialect voices are in fact only one aspect of a range of emerging applications with HMM synthesis, including different types of expressive speech, different speaking styles, or generation of transitions between different speakers (Pucher et al., 2010: 164).

Chapter 12

Summary and further work

We first summarize the main findings of the seven experiments and present a cross-experimental discussion regarding the use of VocalTractLab for articulatory-phonetic research (Section 12.1). A confirmation of the main technical findings can be found in a series of further experiments which we conducted, using VTL for non-articulatory but linguistically oriented studies (Section 12.2). Finally, we discuss some changes that have been implemented in VTL and recently released by Peter Birkholz as VTL 2.1, to put them into relation to our empirical results (Section 12.3).

12.1 Summary and discussion of the experiments

The topics of the seven experiments were quite diverse, to cover different areas of speech production and technical assessment of VTL. The phonetic focus of the speech simulation experiments was located on paralinguistic phenomena of spoken language, and the research questions primarily addressed articulatory details of these phenomena. We suggested articulatory simulation schemata that were then processed with VocalTractLab and subsequently evaluated by acoustic, articulatory and perceptual means. The technical assessment of the modules of the synthesizer encompassed both different anatomical areas of the simulated speech tract and different aspects of the synthesis procedure, especially its robustness and the effects of individual synthesis parameters. Through this evaluation we gained an insight into how practicable VTL is as a research tool, and how adequate its behavior is regarding articulatory details.

The main findings are summarized in Section 12.1.1. Afterwards, we provide an assessment on a more general level, across the individual experiments (Section 12.1.2).

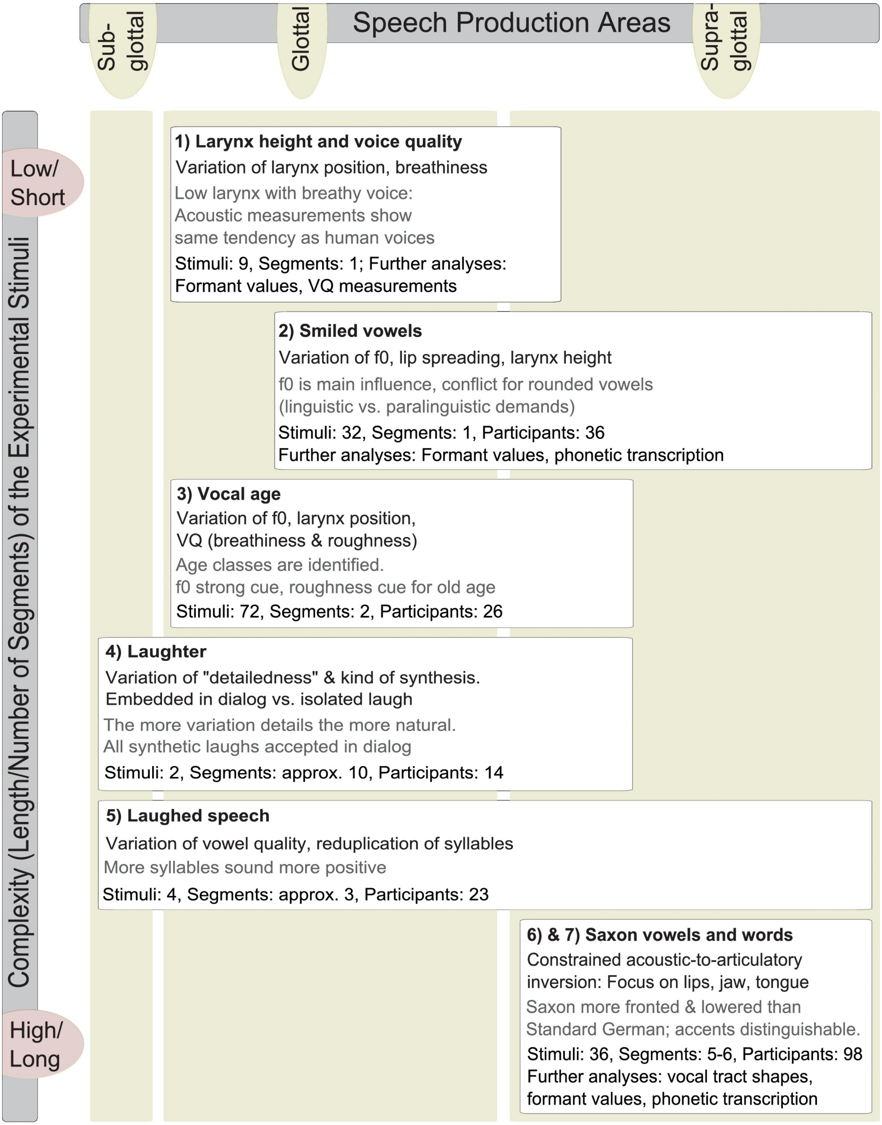
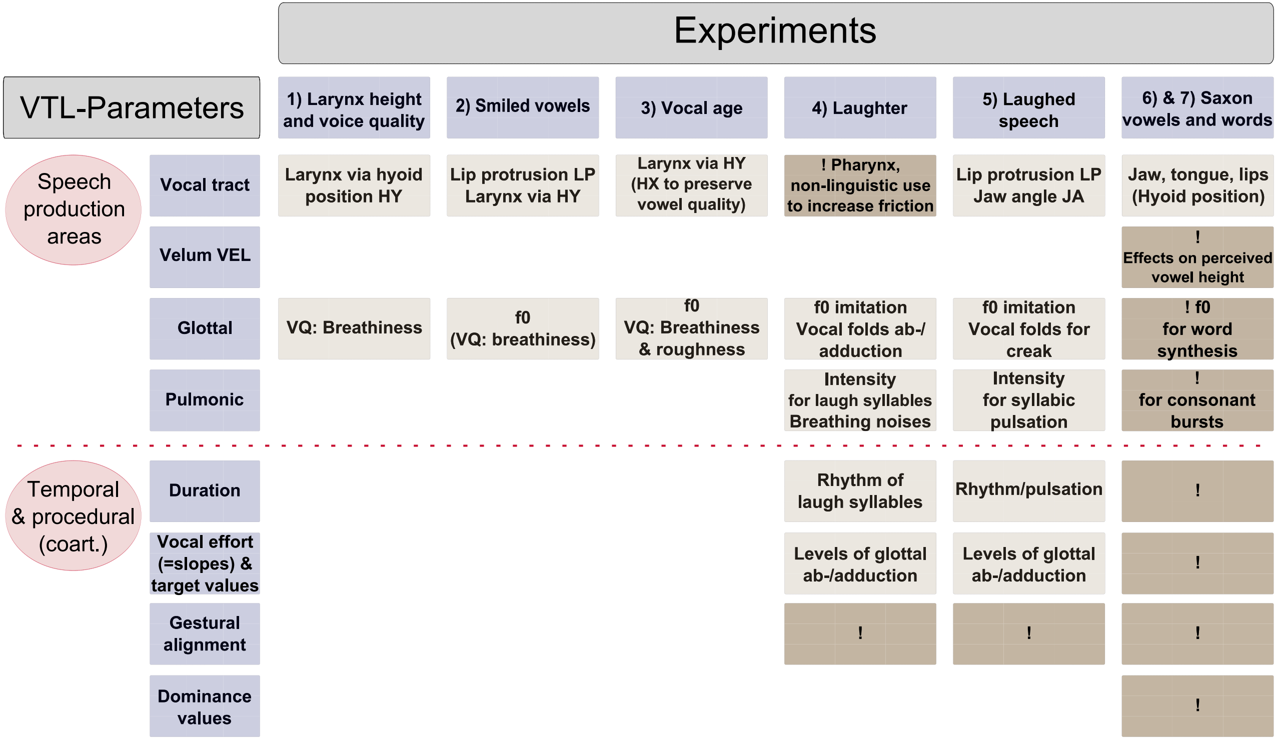

Speech Production Areas
Sub-glottal
Glottal
Supra-glottal
Complexity (Length/Number of Segments) of the Experimental Stimuli
Low/ Short
High/ Long
1) Larynx height and voice quality
Variation of larynx position, breathiness
Low larynx with breathy voice:
Acoustic measurements show
same tendency as human voices
Stimuli: 9, Segments: 1; Further analyses:
Formant values, VQ measurements
2) Smiled vowels
Variation of f0, lip spreading, larynx height
f0 is main influence, conflict for rounded vowels
(linguistic vs. paralinguistic demands)
Stimuli: 32, Segments: 1, Participants: 36
Further analyses: Formant values, phonetic transcription
3) Vocal age
Variation of f0, larynx position,
VQ (breathiness & roughness)
Age classes are identified.
f0 strong cue, roughness cue for old age
Stimuli: 72, Segments: 2, Participants: 26
4) Laughter
Variation of "detailedness" & kind of synthesis.
Embedded in dialog vs. isolated laugh
The more variation details the more natural.
All synthetic laughs accepted in dialog
Stimuli: 2, Segments: approx. 10, Participants: 14
5) Laughed speech
Variation of vowel quality, reduplication of syllables
More syllables sound more positive
Stimuli: 4, Segments: approx. 3, Participants: 23
6) & 7) Saxon vowels and words
Constrained acoustic-to-articulatory
inversion: Focus on lips, jaw, tongue
Saxon more fronted & lowered than
Standard German; accents distinguishable.
Stimuli: 36, Segments: 5-6, Participants: 98
Further analyses: vocal tract shapes,
formant values, phonetic transcription

Figure 12.1: Overview of the experiments, vertically arranged by rising complexity of the stimuli. The length of a box represents the extent to which the respective speech production areas have been covered. Each box shows the topic, test parameters, results, and scope of the empirical setup. *VQ* = voice quality.



		Experiments					
	VTL-Parameters	1) Larynx height and voice quality	2) Smiled vowels	3) Vocal age	4) Laughter	5) Laughed speech	6) & 7) Saxon vowels and words
Speech production areas	Vocal tract	Larynx via hyoid position HY	Lip protrusion LP Larynx via HY	Larynx via HY (HX to preserve vowel quality)	! Pharynx, non-linguistic use to increase friction	Lip protrusion LP Jaw angle JA	Jaw, tongue, lips (Hyoid position)
	Velum VEL						! Effects on perceived vowel height
	Glottal	VQ: Breathiness	f0 (VQ: breathiness)	f0 VQ: Breathiness & roughness	f0 imitation Vocal folds ab-/adduction	f0 imitation Vocal folds for creak	! f0 for word synthesis
	Pulmonic				Intensity for laugh syllables Breathing noises	Intensity for syllabic pulsation	! for consonant bursts
Temporal & procedural (coart.)	Duration				Rhythm of laugh syllables	Rhythm/pulsation	!
	Vocal effort (=slopes) & target values				Levels of glottal ab-/adduction	Levels of glottal ab-/adduction	!
	Gestural alignment				!	!	!
	Dominance values						!

Figure 12.2: Parameters manipulated in each one of the experiments. The experiments are arranged horizontally with respect to increasing complexity of the stimuli involved. Vertically, groups of technical parameters are shown: Speech production area parameters were important in all experiments, while temporal and procedural parameters were mainly important in the experiments with the more complex stimuli. Parameters shown in light-colored boxes were by design in the focus of the experiments, those in parentheses were peripherally manipulated. Parameters shown in dark boxes, with an exclamation mark, were unexpectedly difficult and demanded additional fine-tuning in order to successfully generate the stimuli. *VQ* = voice quality, *coart.* = coarticulation.

12.1.1 Phonetic findings and technical assessments in the experiments

Figure 12.1 shows an overview of the experiments, including a summary of the main phonetic results. Figure 12.2 shows a summary of the different technical parameters focused on in the experiments. As has been discussed in several experiments, some parameters, which were not originally in the focus of an experiment, were unexpectedly difficult to handle (most notably in Experiments 6 and 7). They are marked with an exclamation mark ("!") and detailed further below. The phonetic and technical findings are summarized in the following.

In Experiment I, we manipulated larynx height and voice quality in vowels and were able to confirm general effects that larynx height has on voice quality, especially with respect to breathiness. The permuted combination of all parameters showed 'human-like' acoustic voice characteristics only in the assumed 'human-like' articulatory setting, indicating that lowered-larynx voice is laxer and accompanied by more breathiness than neutral or raised larynx settings.

This experiment showed that VocalTractLab has the capacity to simulate subtle influences of vocal tract shape and excitation quality. Vocal tract shape was altered in vocal tract length by moving the larynx (or a larynx-related parameter, HY) up and down, which showed effects on formant structure as expected. A manipulation of glottal parameters to vary the degree of breathiness likewise showed expected effects on acoustic quality. The independent manipulation of the articulatory sub-systems of VTL can thus be used to test hypotheses about speech production. However, care has to be taken not to produce non-human-like articulatory configurations since the vocal tract does not implement typical coarticulatory co-dependencies of the individual articulators.

In Experiment II, we were able to induce the perception of 'smileyness' in four different vowels by manipulating the phonetic parameters of lip spreading, larynx height, and the level of fundamental frequency. Each one of the three parameters contributed to the perception of 'smileyness' but the effects differed in the different vowels. This was mainly due to the basic articulatory configuration of each vowel, i.e. the /iː/ did not benefit from lip spreading because of the inherently spread lips to start with. For the rounded vowel /uː/, lip spreading seemingly destroyed the basic vowel quality perception. This indicates that a more fine-grained articulatory scheme is necessary to reliably produce 'smileyness'. This scheme, however, is dependent on the technical possibilities offered by the synthesizer.

Technically, we varied horizontal lip spreading (LP), larynx height (HY), and fundamental frequency (f_0) to obtain the smiled vowels. Acoustic analyses showed that all these parameter manipulations induced the theoretically expected acoustic consequences. However, the simple application of lip spreading to the rounded vowel /uː/ obviously altered the timbre too much for it to be recognized as an /uː/ phoneme. Therefore, it seems useful to make an additional lip parameter available in VTL, which could be responsible for pressing the lips together near the corners of the mouth while retracting them.

In Experiment III, we proposed a complex arrangement of vocal qualities associated with three different age groups (YOUNG, ADULT, SENIOR). The manipulated phonetic parameters included fundamental frequency, a breathiness component, and a component of different jitter and shimmer characteristics (‘roughness’). The SENIOR voices were reliably recognized, probably based on their ‘rough’ voice quality, while all other voices were predominantly recognized as ADULT voices. Therefore, the proposed scheme seems to successfully imitate SENIOR voices, while YOUNG and ADULT voices are hard to distinguish. The strongest cue for the age class decisions seemed to be the level of fundamental frequency.

Technically, the articulatory features included vocal fold displacement, usage of glottal leaks, vertical phase lag of the vocal folds, and larynx height. Besides articulatory features, direct signal-related features were manipulated by defining equations for jitter and shimmer behavior. The articulatory synthesis system was therefore used in a ‘hybrid’ way (articulation vs. signal surface), manipulating the usual articulatory parameters which accordingly surface in the acoustic simulation, but also manipulating f_0 and lung-pressure contours *directly*, without defining the relevant articulatory details that would produce these higher-level specifications. This was only possible because VTL is set up in a way that allows fundamental frequency and pulmonic pressure to be controlled directly by specifying their target values (Hz and kPa, respectively). Overall, all these manipulations were only made possible by employing a custom-built batch tool which controlled VTL through a programming interface. Moreover, this batch tool enabled us to synthesize and evaluate a large number of explorative parameter settings to find appropriate value ranges for each age-related parameter.

In Experiment IV, we imitated the complex structure of a song-like laugh, also including breathing noise at its beginning and end. The laugh was accepted by listeners as a seemingly natural laugh in conversation. When evaluating the laugh in isolation, perception ratings showed that higher internal variation during the sequence of laugh syllables increased perceived naturalness. The internal variation was obtained by varying the phonetic parameters duration, intensity and fundamental frequency. Overall, the proposed scheme, which focused strongly on glottal and subglottal activity, seemed appropriate as an initial strategy to simulate an entire laugh from scratch.

Technically, we found that VTL was able to adequately execute the basic articulatory patterns needed for that laugh. This was particularly remarkable since VTL was originally designed for the demands of ‘regular’ speech, which strongly relies on canonical demands for segmental structure. Nevertheless, extensive glottal and subglottal manipulations were possible that helped to create the typical rhythmic laugh structure. However, these manipulations also reached the limits of the synthesizer, especially with regard to peak lung pressure and direction of air flow. As an ad-hoc method to increase friction noises, a gesture for a slight pharyngeal constriction was put into place. For more adequate laugh imitations, it would be beneficial to have access to higher maximal lung pressure and to be able to not only simulate egressive but also ingressive air flow to properly imitate inhalations. Furthermore, it became clear that the gestural alignment is more sensitive to the coordination of a complex arrangement of commands than we previously thought.

In Experiment V, we imitated a short speech-laugh and investigated the perceptual effects of syllabic pulsation and smiled laugh-vowel quality on the degree of perceived amusement. While syllabic pulsation in our speech-laugh stimuli seemed to have a slight effect, none was found for smiled vowel quality. However, the degree of manipulation of vowel quality in our stimuli was rather subtle, so there is no reason to exclude smiled-vowel quality from the speech-laughing system.

Technically, the articulatory parameters were a combination of the ones used in Experiment II (smiled vowels) and Experiment IV (the pulsating structure of laugh syllables), plus the requirements of 'regular' segmental articulation. This complex usage of VTL proved to be feasible in a straightforward manner, mainly because its parameters are defined articulatorily and they can be used in a transparent and flexible way to obtain the desired laugh and speech components of a speech-laugh. Again, the details of the alignment of the gestures proved to be more complex and context-sensitive than expected.

In Experiment VI, we created six long high and mid-high synthetic vowels with an intended Saxon pronunciation by adjusting their articulation so that they would render close acoustic imitations of human vowel pronunciations. We checked the articulatory suggestions visually for plausibility, and then analyzed the articulatory patterns of the Saxon vowels in comparison with their equivalent Standard High German vowel partners. We found a general tendency in our vowel set for a fronted and lowered articulatory setting in Saxon when compared to Standard High German.

Technically, this experiment was mainly driven by the capabilities of a formant optimization function in VTL. Essentially, this is a constrained acoustic-to-articulatory inversion algorithm. It outperforms manual adjustments of the supraglottal parameters when attempting to create a particular vowel sound based on manual articulatory shifts. In combination with a visual plausibility check, it seems to be a reliable tool to acquire articulatory data for the entire vocal tract in a fast and comprehensive manner.

In Experiment VII, the Saxon and Standard vowel sets from Experiment VI were embedded in two-syllable carrier words and subjected to a formal perceptual evaluation. Results indicated an overall recognition of the intended varieties, i.e. Saxon and Standard High German. This supports the results regarding the suggested articulatory shift found in Experiment VI. Furthermore, this experiment illustrated how fine articulatory detail in word pronunciation can reliably induce changing percepts in listeners.

Technically, the experiment made extensive use of word synthesis in VTL by creating a larger number of two-syllable words that were systematically manipulated in their stressed vowel quality. This experiment showed that despite the manual configuration of the gestural scores VTL can be viably used to create entire words with strictly defined articulatory features. However, it also became clear that the current quality of word synthesis is struggling with a relatively low level of 'robustness' against coarticulatory effects, resulting in challenges regarding segmental intelligibility. They can be counteracted by meticulous adjustments of gestural alignment, vocal effort, and dominance values. However, some refinements in the acoustic simulation component of VTL seem necessary to e.g. improve the segmental quality of plosive bursts. The question remains open whether improvement of segmental intelligibility is a matter of optimizing *a)* gestural timings and other attributes of the gestural commands, or *b)* the nature of the simulation and synthesis models. Presumably, it is a blend of the two.

Although the thesis focuses on paralinguistic aspects of speech, we include the assessment of word synthesis (Experiment VII) because a functioning basic linguistic utterance is the prerequisite to applying paralinguistic changes to it, and it is important to have some insight into how the different domains work in VTL. The experiment showed how closely they interact and that it is usually *not* just a matter of simply superimposing paralinguistic features onto the segmental content because an exchange of vowels in the accented words often affected coarticulatory and neighboring segmental quality as well. Since similar kinds of coarticulatory influence are also important in regular speech ('long domain' coarticulation, cf. e.g. Heid and Hawkins, 2000), it seems all the more critical to take care of them during synthesis.

12.1.2 Discussion across experiments

We discuss the main strengths of and challenges in VocalTractLab as they have been showing across our articulatory-phonetic studies. The issues are mostly technical in nature, although at the end of the section we discuss two issues connected with the experimental design.

Strengths

VTL has been remarkably flexible regarding the paralinguistic demands that it successfully simulated and can be considered a useful tool for many speech tasks. It became clear in the series of experiments that the basic requirements for articulatory research are met by VocalTractLab, and it proved to be a valuable tool for basic phonetic research although it is not yet able to perform text-to-speech synthesis. Its strengths lie in providing a simulation of the entire speech apparatus within a single coordinate system, in making transparent the possible articulatory movements and settings that underlie a desired audio output, and in providing direct control of all parameters individually. The completeness of the system combined with the transparent nature of its parameters enables us to design systematic series of experiments regarding the relationship between articulation and acoustics.

Compared to other types of synthesis, articulatory synthesis may not have the inherent naturalness of the voice. However, it offers high flexibility in designing the speech output and requires no post-hoc signal processing. This strength can be applied especially well when creating expressive speech or sounds such as laughs, speech laughs, and breathing noises.

With the provision of the formant optimization algorithm used in Experiment VI for Saxon vowels, VTL offers a useful tool to create paralinguistically marked phones similar to the base phones in the default phone set. However, since it optimizes the articulation via a match in the acoustic domain, we diverge slightly from the articulatory principle. On the other hand, since acoustics in general cannot be mapped one-to-one onto articulation, this approach represents a pragmatic way to handle articulatory underspecification and gets us closer to the 'discovery procedure' goal. The great benefit of the algorithm lies in the fact that it enables us to create sounds that match very specific acoustic demands. To obtain this matching turned out not to be feasible by manual adjustments of the articulators because the applied movement always tended to be too coarse.

Challenges

During the experiments, some challenges have become obvious. Therefore, to ensure articulatory adequacy and high-quality utterance output, several technical aspects have to be paid attention to. First, as became clear e.g. in the experiment on larynx height and voice quality (Chapter 5), one has to be careful to only select naturally attainable articulatory settings, and as became clear in the experiment on Saxon-accented vowels (Chapter 10), one has to pay special attention to human-like coarticulatory co-dependencies of e.g. the tongue and the jaw: In default human articulation their movements are closely inter-connected, but this is currently not implemented in VTL. Of course, humans are also able to move the tongue and the jaw separately. However, we think one should distinguish between movements that are possible and those that happen regularly during speech. VTL itself currently has no 'restrictive layer' to prevent unnatural shapes of the vocal tract and comprehensive modeling of coarticulatory co-dependencies is currently not available.

Second, especially for word synthesis but also for all other utterance types, it would be convenient to have a more 'robust' handling of transitions and coarticulation across neighboring segments, be it in consonant-vowel sequences or in consonant clusters. Currently we have to pay attention to the adequate gestural implementation of fine articulatory detail, such as the inter-dependencies of the vocal effort of one segment and the duration, quality and perceptual prominence of a neighboring segment. Similarly, the elevated context-sensitive performance of a number of consonants requires additional individual adjustments in phone definitions, as was found in word pairs such as /zuːmən/ <zoomen> 'to zoom', which sounds alright, vs. /ziːmən/ (non-sense word), which produces a lisp in the fricative. VTL, however, does offer a number of phones in the standard phone set that are specifically optimized for certain contexts, such as the phone entries "g(i:)", "g(a:)", "g(u:)." As far as we know, they served as initial and preliminary context-sensitive configurations during software development, from which a single, more general one was derived, i.e. a general "g" as a merger of "g(i:)", "g(a:)", "g(u:)." Perhaps it is an option to use the initial configurations regularly for word synthesis.

Third, the context-sensitivity leads to the most salient technical challenge in VTL, namely the extensive effort needed to create gestural scores for any unit longer than a few sounds. Therefore, within the framework of this thesis we did not expand the complexity of the stimuli beyond the word level or very short phrases. Additionally, and perhaps due to the high effort needed to achieve a good pronunciation, it seems that word synthesis quality depends greatly on the words chosen, or, more precisely, on the segments comprising them. This is why we find a strong influence of carrier words in the Saxon accented words in Experiment VII. A higher level of coarticulatory robustness would help in the creation of gestural scores by reducing the need for 'tweaking' them, because the minute differences between scores would not affect the outcome so much if some kind of coarticulatory robustness was present. Currently, the use of the rudimentary TTS functionality provided by the song file format is a way to speed up the process of creating words or phrases since it eases the creation of basic gestural scores. This illustrates how, by implementing rules, it is possible to provide a rough first approximation of the articulation of a given word but it certainly has to be developed further in order to make VTL suitable for more large-scale applications and to provide full TTS capability.

Fourth, in some areas we reached the limits of currently available ranges of parameters such as pulmonic pressure for laughter, or lip shape control for smiled speech. For adequate and high-quality synthesis, broader ranges and additional parameters are required.

Fifth, to make the simulation even more naturalistic, ingressive air flow and ingressive speech capability in general would be valuable. Ingressive speech is more widespread than one might think when considering the established properties of current synthesis systems. This feature could therefore contribute to increasing the naturalness of the synthesized voice, e.g. when applied in dialog systems, because ingressive speech is used as a backchannel utterance or for interjections in many languages (Eklund, 2002, 2007). Interestingly this has been noted to occur only in human-human communicative settings, not in human-machine settings. Perhaps, if machines used ingressive voice themselves, human behavior might change, indicating that the machine is accepted as a more naturalistic conversational partner.

Finally, two issues regarding technical limits are addressed that had an impact on the empirical designs. Firstly, we did not test for gender specific effects of any kind in the experiments. It is not possible to create an acceptable female voice with the present synthesis system and we believe this should be addressed in future work. This concerns articulatory research as well as perceptual evaluation, i.e. simulation experiments should ideally be conducted with a male as well as a female synthetic voice, and during perceptual evaluation the number of participants should be balanced for gender.

Secondly, throughout our experiments we did not explicitly test for intelligibility of the utterances, although during immediate informal evaluation we ensured that basic intelligibility was maintained. Since the intelligibility was often quite low, it seemed expedient to use written words along with the audio stimuli. Thus, a systematic testing of intelligibility was not in the focus of our work. However, since paralinguistic features of voices and utterances interact with the segmental linguistic content of the utterances, systematic intelligibility evaluation should ideally be done.

12.2 More general use of VTL in linguistic experiments

The articulatory-phonetic speech production experiments presented so far in this thesis focused primarily on *articulatory* aspects of speech production and on the effects of articulatory differences on speech perception. However, VTL can also be used to study questions in which articulation is not the primary focus of interest. With the knowledge and awareness gained from the previous phonetic experiments, we can create word or phrasal stimuli that can be successfully used for ‘regular’ linguistic experiments. The articulatory details are simply supposed to work in the background and the experiments do not explicitly test the synthesizer. Instead, other research questions are addressed.

This has been done e.g. in a series of studies regarding third-party perception of certainty and uncertainty in a fictitious question-answering scenario between a human and a machine, involving 13 group perception tests with a total of 390 participants (Wollermann and Lasarcyk, 2007; Lasarcyk and Wollermann, 2010; Wollermann et al., 2013; Lasarcyk et al., 2013). In these studies, we created short phrasal stimuli in a way similar to the procedure described in Section 11.1, using the rudimentary TTS function of the song file format and adapting the synthetic words to make them match the pronunciations recorded from a human speaker. Besides ‘regular’ words, the stimuli also featured the filler words or hesitation particles <hmm> [mː] and <äh> [ʔɛː]. They were fine-tuned to suit the phonetic context of the surrounding phrase. Systematically varied stimuli were created by presence or absence of these fillers in the synthesized phrases, by changing the intonation patterns, by inserting different amounts of silence between fillers and content words, and by varying the response delay.

From a synthesis technical point of view, the main outcome confirms the findings and limitations discussed in the thesis experiments. While we were easily able to exploit the suprasegmental level of stimulus variation (e.g. pitch manipulations), the segmental level showed some difficulty, resulting in occasional misunderstandings of the words. For instance, voiced plosives were heard as fricatives – /ˈboːnən/ <Bohnen> ‘beans’ as [ˈvoːnən] <wohnen> ‘to reside’ – and places of articulation were not correctly recognized – /ˈɡʊɐ̯kən/ <Gurken> ‘cucumbers’ as the nonsense words [ˈtʊɐ̯kən] or [ˈdʊɐ̯kən]. This again indicates that word synthesis still needs to be improved on the segmental level (cf. Section 12.3.4).

Nonetheless, this series of studies illustrates that the synthesizer is in general suited to conducting analysis-by-synthesis perception experiments to test effects of different speech and language[1] parameters. Synthetic speech is used here to test *high-level* factors of spoken communication rather than ‘low-level’ articulatory features as addressed in the thesis experiments. VTL offers a method of synthesis which allows very precisely tailored stimuli to be constructed that are able to also meet potentially extreme demands such as high f_0, without introducing the synthesis artifacts common in other synthesis methods.

[1] Language in the sense of presence or absence of a word and not so much of how this word is uttered in detail.

For ‘regular’ linguistic experiments, other synthesis methods are currently more usual, and the creation of a basic stimulus set may be more time consuming with VTL than when using formant, concatenative, or HMM synthesis. Moreover, up to now, the complexity of the synthesized speech material has its limitations, especially if phenomena such as consonant clusters are involved. However, VocalTractLab can *also* be used, and its strength lies in the capacity to easily incorporate atypical utterances and non-words in stimulus materials. Additionally, the typical caveats of articulatory synthesis, such as consonant coarticulation or plosive bursts, are subject to intensive ongoing research. Therefore articulatory synthesis, or VTL, can gradually be considered a general alternative synthesis technique in the near future.

12.3 Further development of VTL: VTL 2.1

At this point, we think it is appropriate to note how the synthesis system has recently been developed further (Birkholz, 2013a,c). We briefly highlight some main changes in the newest version (VTL 2.1, P. Birkholz, pers. comm.), focusing on aspects that have been addressed within the thesis, and relating how the advances might affect the findings of our experiments.

12.3.1 Voice quality and new vocal fold models

In VTL 2.1, it is now possible to directly select particular voice qualities from predefined profiles. Additionally, one can choose between different vocal fold models (Birkholz and Neuschaefer-Rube, 2012). Each model is designed to cover the full range of voice qualities, but they differ in which quality is produced best. The overall evaluation is ‘good’ for all models, therefore they are sufficient for regular synthesis purposes (but not as sophisticated as models such as the one described in Moisik and Esling, 2007; Moisik, 2008). Among the glottis implementations we find the model from Titze (1984), the classical self-oscillating two-mass model by Ishizaka and Flanagan (1972), and the modified two-mass model by Birkholz et al. (2011b). They produce different sounds due to their different definitions, which e.g. influence the resulting VOT in consonant-vowel sequences. All of them have two control parameters in common (f_0, p_{sub}) which are directly controlled in the gestural score.

The new profiles and models may make the voice quality manipulations related to larynx height (Chapter 5) or aged voices (Chapter 7) somewhat easier but the aim of the experiments was to freely manipulate voice quality parameters instead of using predefined profiles.

12.3.2 Changes in vocal tract parameters

A small number of the vocal tract parameters have been changed in VTL 2.1 to more realistically map the scope of possible movements of each articulator. In particular, the jaw now has only two instead of three parameters; the velum is now controlled by two parameters instead of one (Birkholz, 2013a).

These changes may have an impact on results such as the modeling of Saxon vowels (Chapter 10) since the formant optimization algorithm would work on slightly different parameters. The essence of the modeling should not be substantially affected though.

12.3.3 A new model of coarticulation

As has been addressed at several points in this thesis, the current implementation of the dominance model in VTL (see e.g. Birkholz and Kröger, 2006) leads to some unwanted coarticulation phenomena since it does not yet properly take into account the segmental context. Recently, a new approach has been implemented which uses a new kind of interpolation model. It is more precise than the old one, in particular it features more than one vocal tract shape for each consonant, depending on the vowel context. Now three profiles are always defined, for the vowel contexts /i u a/, e.g. /bi bu ba/. If an utterance contains the sequence /bɛ/, the model interpolates between /bi/ and /ba/.

This method of context-dependent definition and subsequent interpolation would probably make the word synthesis in Experiment VII (Chapter 11) more robust. For the time being, we used a preliminary solution with individual context-sensitive phone definitions, therefore the results should not differ greatly. However, the synthesis process should now be more direct and straightforward since the previously desired context-sensitivity is now already implemented in the software.

12.3.4 Simulation of segmental targets

The segmental intelligibility of VTL will be considerably improved in VTL 2.1, especially for obstruents. Currently, a new technique is developed for a 3D real-time reconstruction of the mouth cavity. Along with recordings of aerodynamic parameters and acoustics, it will be used to improve noise source modeling in VTL (Preuß et al., 2013). Once this is achieved, the overall quality of word synthesis should improve by reducing misrecognitions of consonants.

Combined with the new coarticulation model, segmental quality should become more robust. This may support the ease of stimulus creation in word synthesis such as in Chapter 11. It may also positively affect perceptual ratings of such stimuli due to better segmental intelligibility.

In summary, the sketched changes in VTL 2.1 should have a positive impact on the work flow and ease of achieving good sounding utterances while not substantially changing the results of our articulatory experiments if they were to be re-run using VTL 2.1. Some of the changes that are now systematically implemented have already been used in a comparable way as ad-hoc solutions in our experiments, such as the context-sensitive dominance modeling in phones. One characteristic strength of VocalTractLab, the freedom to manipulate all kinds of articulatory parameters individually, will still be present in VTL 2.1 as much as in VTL.

Chapter 13

Conclusions

We conclude by reviewing the basic idea of combining articulatory-phonetic research with articulatory speech synthesis (Section 13.1) before presenting a brief outlook on possible future work (Section 13.2).

13.1 Articulatory-phonetic research and speech synthesis

The work carried out for this thesis has illustrated how state-of-the-art articulatory speech synthesis can be used for articulatory-phonetic research. A series of articulatory-phonetic experiments have been conducted that were designed *a)* to investigate fine phonetic details of different paralinguistic properties of the human voice, and *b)* to evaluate different components of the articulatory speech synthesizer VocalTractLab by covering the different areas of the articulatory speech apparatus. The phonetic findings suggest new articulatory details that extend the understanding of the production of traits of smiling in the voice, laughter, aging, and regional accent. The findings of the technical evaluation of the synthesizer illustrate the current state of VTL as a phonetic research tool, covering issues such as vocal tract configurations and coarticulation, gestural score organization, and fidelity of the acoustic simulation with regard to fine articulatory detail.

The motivation for using an articulatory synthesizer such as VTL for articulation research was based on the view that such a synthesizer represents a software framework which, in a sophisticated way, bundles relevant findings and models of speech production in a single comprehensive system. It covers the whole speech tract and is equipped with an acoustic module that permits direct acoustic evaluation of articulatory events. Therefore, it offers a direct insight into details of articulation and their effects on acoustics and perception.[1]

VTL produces speech output that is of high quality, as has been demonstrated in the demo material submitted together with Peter Birkholz's doctoral thesis (Birkholz, 2006), by

[1] This feature also makes VTL very powerful when applied as an educational device. The interactive articulation manipulation, which can be listened to immediately, provides valuable material for teaching the basic articulatory-phonetic and acoustic inter-relations.

using parameters which are articulatorily grounded and comprehensible to the user. This was taken as an indication that the system really puts together, in a sensible way, existing knowledge from speech production and acoustics. Known simplifications and limitations, which are part of every model, were taken into account when interpreting the articulatory results of the simulations.

All in all, the combination of phonetic basic research and articulatory speech synthesis is one of mutual influence and evaluation. This has already been mentioned in the thesis Introduction and will be discussed further in the following.

As far as the first direction of influence is concerned, VTL can be used as an evaluator of articulatory and physiological models to further articulatory understanding on the level of fine articulatory detail. By evaluating the articulation mechanisms at the 'edge' of spoken communication, i.e. paralinguistic features, it can uncover the limits of current articulatory models and extend our present phonetic knowledge. Using VTL as an evaluator shows that our present knowledge of articulation – and of aerodynamics or acoustics – and the models that are based on this knowledge are not yet comprehensive enough for a complete model of articulatory-acoustic simulation.

The strength of VTL as an evaluator lies in the opportunity to investigate articulation within a complex model without having to deploy intrusive and possibly stress-inducing tools for physiological-articulatory data acquisition. On top of that, we gain the opportunity to develop and test research questions that can be defined on a very precise level of detail since the investigated articulation movements take place within a single technical system using well defined parameters within one coherent spatial representation. In this way we can test specific aspects of articulation more precisely and in a more controlled manner than by the direct analysis of human recordings.

It holds without a doubt that instrumental articulatory research performs invaluable groundwork. It studies the individual aspects of articulation with different instrumental techniques, mostly in locally focused areas, often having to simplify certain aspects, and seeing more questions arise than are answered. Nonetheless, these contributions produce the valuable pieces of the puzzle which are then brought together by such systems as VocalTractLab. These two lines of research therefore complement each other.

As far as the second direction of influence is concerned, the experience gained in the phonetic simulation experiments can support the ongoing development of articulatory synthesis and contribute insights to help it become more robust and still more accurate because the conscious and targeted application of the synthesizer to particular pronunciation tasks has brought to light the current strengths and limitations of the system (Section 12.1.2).

13.2 Outlook

The experiments in this thesis spanned a two-dimensional area of investigation and evaluation, moving along different paralinguistic simulation tasks and technical parameters (see e.g. Figure 12.1 or 12.2). Thanks to the diversity of evaluation issues, the series of experiments already gives a comprehensive insight into the status of VocalTractLab as a research tool and contributes a number of aspects to the understanding of fine articulatory detail in speech. Nevertheless, studying the following issues could build on and complement the previous experiments.

With respect to the synthesizer's components and technical parameters one could evaluate the following: Expanding on the measurements of voice quality (Chapter 5), one could measure the acoustic properties of different nasality settings and evaluate the settings perceptually. This could clarify questions that were raised during the accented vowel simulation (Chapter 10), as to how the degree of velic opening should be adjusted to approach near-human speech characteristics in vowels of different vowel height.

Since word synthesis was only tested in the final experiment, this area also needs to be pursued further to complement the previous experiments. Investigations could start with systematic testing of consonant-vowel sequences, which are in a sense the basic building blocks of words, and features such as VOT, which are controlled by details on the gestural score. This could give insights into permissible ranges of duration, amplitude, and temporal location of glottal (aspiration) gestures for CV sequences (cf. also ongoing work on VTL presented in Birkholz, 2013a). When the knowledge of robust 'low-level articulation' is available, it can be combined with automatic prediction of higher level features to, at some point in the future, constitute a fully automatic TTS system.

With respect to paralinguistic features, the phenomena to test are innumerable. We therefore propose a phonetically motivated grid on a meta-level to find sensible extensions to our series of experiments. The grid covers five important phonetic dimensions: voice quality, spectral structure (timbre), intensity, pitch (f_0) and duration. The relatively limited degree to which these dimensions have been addressed in this thesis suggests the scope for research that remains open.

The first one, voice quality, has been addressed e.g. in Chapter 5, by studying interactions of voice quality and larynx height, and in Chapter 7, by using voice quality as a property for aged voices signaling speakers of different ages. But voice quality is also communicatively important in other ways, for example as a signal of speaker health and mood or status.

The second dimension, spectral structure, has been addressed e.g. in Chapter 10, by illustrating interactions of formant characteristics based on a fronting and lowering of vowels and perceived accent. There is clearly more to be done here to address the old question of *articulation basis*, not least in pursuing the question of parallel effects on consonants of the articulatory differences between accents (e.g. fronting and lowering) that were found for vowels.

The last three dimensions, intensity, pitch and duration have only been touched on in this thesis, in cases when we aimed to produce close-copy imitations of human recordings. This was the case e.g. when simulating the laugh (Chapter 8) in which intensity, duration and f_0 of the laugh syllables were tuned extensively to match the human laugh.

Technically, the investigation of intensity and pitch as features in paralinguistic phenomena would be relatively straightforward, namely by adjusting pulmonic pressure and f_0 values. The interesting part is how they interact with other dimensions: Combined with different voice qualities one could e.g. investigate the attributes used to establish privacy in conversation. Combined with articulatory precision, one could investigate properties of yelled speech.

The dimension of (segmental) duration has not been addressed systematically because of its close relation to the perceived rhythm of speech, which becomes important in word synthesis and even more important in the synthesis of running speech. Rhythm was, by design, not a topic in itself in the previous experiments. However, we have addressed general aspects of the control of segmental duration, pointing out that segmental duration is not only a matter of gestural duration but also of the velocity and the amplitude of a gesture and its overlap with neighboring gestures.

Segmental duration can therefore be linked to the question of segmental modification in regular, casual and fast speech, a problem for which articulatory synthesis seems particularly well suited. Since it is relatively easy to compress the time scale of the gestural score, one could e.g. study reductions as they occur in rapid speech by generating an increasingly large overlap of commands in unstressed syllables. As a consequence, segments become acoustically hidden but are not elided on the articulatory level (cf. e.g. Browman and Goldstein, 1989, 1990). This would be a blend of the paralinguistically motivated dimension of articulation rate as a speaking style and the phonological question of the existence of (acoustically covert) gestures. The current challenge in testing fast speech with VTL, however, is that the word synthesis quality even at regular speaking rate is relatively low.

Finally, a general complement to the experiments reported in this thesis could be to conduct systematic coarticulation experiments, such as coarticulation across syllables or ‘long domain’ coarticulation as has been addressed e.g by Heid and Hawkins (2000). Such investigations were by design not part of our experiments because we primarily investigated paralinguistic features.

All in all, we hope to have illustrated three things in this thesis: Firstly, by studying paralinguistic phenomena of the voice we hope to have contributed some aspects of knowledge about fine articulatory detail that may extend our understanding of articulation. Secondly, we hope to have illustrated how VocalTractLab performs when it is employed as an articulatory-phonetic research tool, exemplifying what current state-of-the-art articulatory synthesis is capable of accomplishing. Lastly, in bringing the two strands together, we hope to have contributed a puzzle piece to one day creating naturalistic sounding, flexible synthesis that is not unduly challenged by the demands of expressive speech, as most systems are nowadays. This type of synthesizer could be part of TTS systems that are able to produce expressive speech with a voice that has individualized features such as age and accent, without having to prepare and hold available large databases of recorded human voices.

Appendix A

Acoustic settings of VTL and their impact on vowel quality[1]

This part of the Appendix provides background information on the two synthesis modes that can be used in VocalTractLab (Section A.1). We illustrate how they can influence the acoustic outcome of synthesized vowels and that it is therefore important to document the acoustic settings used in an articulatory synthesizer (Section A.2).

A.1 Synthesis modes

During the process of aerodynamic-acoustic simulation in the articulatory synthesizer VocalTractLab (VTL, Birkholz, 2006), a geometric vocal tract target configuration is used as the general basis to generate a sound. A discrete area function is derived from this geometry, upon which the actual aerodynamic-acoustic simulation is calculated. It is hard to obtain an optimal speech outcome due to constraints e.g. in the numeric simulations that have to be applied to simulate all the necessary acoustic effects. There does not seem to be a consensus about how to deal with all the relevant factors (cf. Sundberg et al., 1992; Wakita and Fant, 1978; Badin and Fant, 1984). Such factors can be e.g. the degree of nasal coupling, the kind of voice excitation, the losses that are computed, how the radiation impedance is considered, and how subglottal and glottal coupling are integrated.

In VTL, one can choose between two basic simulation methods, frequency-domain simulation (FDS) and time-domain simulation (TDS). Each simulation faces different constraints when simulating the acoustics from a given area function and presents the user with different options on how to render the acoustic signal. The possible acoustic settings in VTL, described in Table A.1, reflect the choices that can be made for each synthesis method. On the one hand, the acoustic settings comprise the core-technical settings for each synthesis

[1] This chapter presents material in a shortened and adapted form, previously published in Lasarcyk (2010).

method (FDS or TDS). On the other hand, they also comprise additional synthesis settings such as velic aperture. From a user's perspective, they appear as an acoustic setting because they are controlled outside of the plain vocal tract target configuration. They may thus silently alter the vocal tract area function used for the acoustic-aerodynamic simulation if the user does not pay explicit attention to these settings.

The important point to note here is that articulatory synthesis can create more than one sound from a single underlying vocal tract target configuration (the theoretical phone description). Firstly, the acoustic settings influence the way the vocal tract target configuration is transferred into a corresponding vocal tract area function, yielding different area functions from one single target configuration. Secondly, one area function can produce different speech signals depending on the acoustic settings selected for the aerodynamic-acoustic simulation. It is thus not completely reliable to state that a particular vowel target configuration *per se* sounds like X, because the sound of the vowel also depends on the strategy of the acoustic synthesis.

The two rightmost columns of Table A.1, listing the default values for each synthesis method in VTL, are called *synthesis profiles* in the following. The FDS profile is used e.g. to generate a single vowel sound directly from a vowel vocal tract target configuration. In default mode, it entails the generation of a high-low (98–120–77 Hz) intonation contour to utter the vowel. The TDS profile is used to generate speech from a gestural score. For each time step of the simulation, the shape of the vocal tract is computed and used for time-domain synthesis. Our default gestural score was programmed with certain default values including velic aperture (0.23), glottal area (0.24) and intonation (level contour at 109 Hz).

The following section and Lasarcyk (2010) further illustrate some auditory impacts that these synthesis profiles can have on listeners who are asked to classify different vowels. In the case study presented here, the choice of synthesis profile mainly altered the perceived vowel height.

A.2 Effects of synthesis mode on vowel quality

During the transcription task for the evaluation of Saxon vs. Standard German vowel pronunciation (see Section 10.3.1), we found that there was a confounding issue regarding the acoustic settings used for the acoustic rendering (synthesis mode) from articulatory vocal tract definitions. This section presents auditory and acoustic evaluation of this issue.

To evaluate the vowels auditorily, we select mainly Standard High German vowels for a separate transcription task. They sound more familiar to German native speakers, regardless of which regional background, and therefore the transcription is expected to be more consistent than one of regional-accented vowels. Only the sound /eː/ is presented in both varieties, for illustration purposes.

Table A.1: Acoustic parameters and their possible values for acoustic synthesis in the synthesis framework VocalTractLab, as they can be found in the settings menus for FDS and TDS settings. Not all parameters are applicable to both synthesis methods (empty cells).

Parameter	**Description**	**Possible values**	**Default for FDS (FDS profile)**	**Default for TDS (TDS profile)**
Glottal excitation	Model used for voiced excitation	FDS: Liljencrants-Fant (LF) model; TDS: Titze model	LF model	Titze model
Glottal area	Degree of glottal abduction	[0, 1]	0.00	0.24
Radiation impedance	Kind of approximation that is selected for radiation impedance	FDS: 0 (ideal soft opening)/Piston in sphere/Piston in wall/Parallel circuit of R and L	Parallel circuit of R and L	

Energy losses	Which kinds of energy losses in the vocal tract tube are considered	Yes/no;
		FDS: Boundary layer resistances; Heat conduction losses, Absorbing/-soft walls; Hagen-Poiseuille resistance;
		TDS: Fluid dynamic losses, Soft walls, Sound radiation from skin
			Boundary layer resistances,
			Absorbing/soft walls
				Fluid dynamic losses,
				Soft walls

Signal resistances	Whether extra small signal resistances are used	FDS: yes/no	Yes	
Nose sinuses	Whether nose sinuses are considered or not	FDS: yes/no	Yes	
Piriform fossa	Coupling of the sinus piriformis to the vocal tract	Yes/no	No	No
Lumped elements	Whether lumped elements in T-sections are considered	FDS: yes/no	Yes	
Velic aperture	Degree of velic opening (nasal coupling)	[0, 1]	Defined in vocal tract target configuration for each vowel (0.00)	Defined in gestural score (0.23 in our case)
Intonation	Contour of the fundamental frequency for an utterance	Variable	High-low (98–120–77 Hz)	Defined in gestural score (constant at 109 Hz in our case)

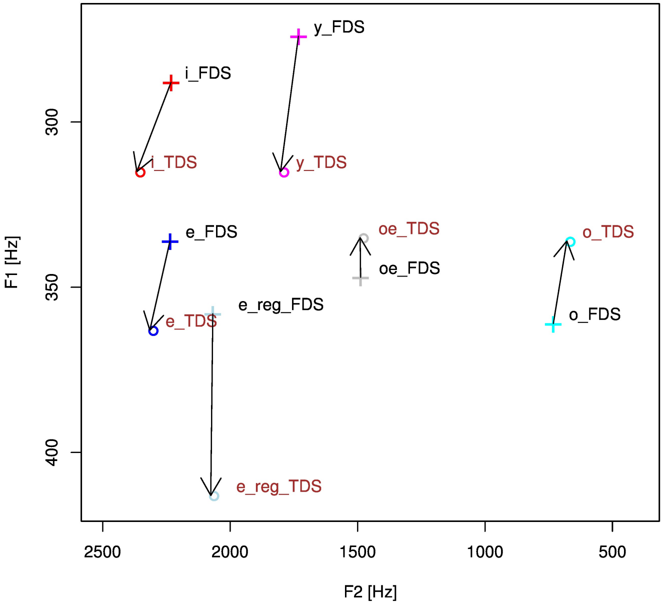


Figure A.1: Sample vowels synthesized using the two different synthesis modes. Each vowel is based on the same predefined vocal tract target configuration and synthesized twice: The first acoustic rendering uses default settings of the frequency-domain synthesis (FDS, black labels), the second one uses default time-domain synthesis settings (TDS, gray labels). Formant frequency listings and phonetic transcriptions are presented in Table A.2.

The transcriptions of each stimulus are presented in Table A.2, along with the formant frequency values. Figure A.1 shows the corresponding F_1-F_2 formant frequency plot, visualizing the differences between the two synthesis modes.

In summary, it becomes obvious that merely due to the technical choice of synthesis mode, critical vowel height differences are perceived. The impact varies for the different vowel qualities. The general tendency in the results is that the acoustic settings used in the TDS mode favor a perception towards more open vowels. Horizontal tongue body position and lip rounding are also influenced, but to a lesser degree and not homogeneously. Additionally, the TDS vowels induce a nasalized impression, mainly due to the increased velic aperture in the TDS profile.

The transcribers choose categorically different symbols to transcribe sounds that originated from one and the same vocal tract target configuration in the default phone setting of VTL. However, differences in vowel height are actually part of the research question of the

Table A.2: Left: *Formant frequency values* of the vowels presented in the transcription task and depicted in the F1-F2 plot in Figure A.1. Values are means taken from the middle 30 % of each vowel, in Hz. Right: *IPA transcriptions* of the two vowel sets, obtained from six trained phoneticians. Slashes indicate two suggestions from the same transcriber. Diacritics used (from top to bottom): centralized [e̽], nasalized [ẽ], more rounded [e̹], less rounded [e̜], retracted [e̠], advanced [e̟], raised [e̝], lowered [e̞].

Vowel	**Set I (FDS)**			**Set II (TDS)**			**Set I (FDS)**	**Set II (TDS)**
	F$_1$	**F$_2$**	**F$_3$**	**F$_1$**	**F$_2$**	**F$_3$**		
/o/	361	732	2517	336	679	2617	o o̞ o̜ o ɔ̟̝ o	ɔ̞ ɔ̠ ɔ̝/ɔ̝̃ ɔ̃ ɔ̃ u̞
/ø/	347	1488	2229	335	1490	2410	ø ø ø̠ ø ø ø	œ œ̝ œ̝ œ̝/œ̜̝ ø̠ ø̞
/y/	274	1732	2214	315	1802	2424	y y ʏ ʏ/ʉ̞ y̽ y̜/y̽	y̜̞ ø̝̽/y̞̽ ø̠ ø ʏ y̠
/i/	288	2232	3069	315	2366	3098	i i̞ ɪ i̠ i̽ i̞/e̝	i e̠̝ e i̞/e̞ e e̝
/e/	336	2236	2799	363	2315	2798	e e e̠ e e/e̝ e	ẽ/ẽ̞ e̝ e ẽ̽ e̠̞ e
/e$_{reg}$/	358	2068	2739	413	2076	2779	e̝/e̝̽ e e̠ e e̹̠̞ e̞/e	ɛ̹̃ ɛ̝ e̹ ɛ̝̃ ɛ̝ e̞

original experimental setup. Thus, we find a critical interference of technical implementation and phonetic investigation. Implications are, firstly, that technical settings in articulatory synthesis should be reported in considerable detail and, secondly, that one should be aware of these kinds of effects because they can confound empirical results.

Appendix B

Web-based listening tests

Since we believe that internet-based listening tests are a valuable but rather new method, this part of the Appendix provides background information on listening tests using the world-wide web, and depicts the design details used for the tests in this thesis. It thus complements the information given in Section 4.2.7. After addressing general characteristics on internet-based testing (Section B.1), we point to measures of quality assurance when designing a test (Section B.2) and when reporting the results (Section B.3), and then briefly discuss implementation issues (Section B.4), followed by screenshots from our experiments (Section B.5).

B.1 Characteristics of web-based testing

The introductory comments and the best practices discussed in the following sections are based on works by Reips and Bosnjak (Reips, 1997, 2002a,b, 2007; Bosnjak, 2003), complemented with our own experiences. Internet-based testing can be regarded as an extension to computer-aided testing in laboratory settings. Its advantages, together with advantages over paper-and-pencil testing, motivated us to use this kind of method for the majority of perception tests. The data are digitally stored at no extra cost, same as in the computer laboratory setting, therefore avoiding mistakes due to manual transfer of data. In contrast to laboratory conditions, no appointments, traveling and continuous personal supervision are necessary, and for some experiments, it is easier to find participants that match special requirements, e.g. with different dialectal backgrounds. The automated, unsupervised process of the test may render it more objective since no individual instruction differences or other non-standardized situations occur.

It can be argued, though, that due to this lack of direct supervision this method is not reliable. However, recent experiences and evaluation indicate that internet-based experiments are nevertheless a sophisticated and reliable method of empirical data collection. This is especially true when a number of rules or best practices are observed for designing a web-based experiment, and when interpreting and reporting the results. These measures can be seen as a strategy for indirect quality assurance and probably help to make this method of data collection (even more) reliable. The unsupervised character of the test situation may be beneficial because participants feel more free to quit a test than producing flippant answers in order to get done with the test more quickly under the eyes of a laboratory supervisor.

B.2 Test design issues

In order to ensure high data quality regarding the design of a web-based test, a number of measures can be taken, four of which are highlighted in this section. The first important aspect is to be able to reconstruct the test environment by e.g. putting questions about it into a questionnaire or logging browser configurations such as display size. This knowledge can serve as indirect, asynchronous supervision to be able to identify potential distractors or problems. Data sets can be excluded if e.g. distractions or noises are reported or if they can be implied from the general test environment (in descriptions such as “internet cafe” vs. “quiet office”). Logging of reaction times can also be an effective tool to detect whether a participant has been losing focus.

The second aspect that can compromise the quality of the data especially for *speech* perception tasks is the fact that every participant uses different audio output devices at their computers, such as head-phones or loud speakers, which all may output different sound qualities. This is added to the generally existing problem of potential different hearing capabilities among the participants themselves. To obtain a handle onto these issues we included questions in the questionnaire about hearing problems and what kind of audio output device was used. We asked the participants to use head phones but did not exclude anyone if they reported of not having used them.

The third aspect we paid attention to was to reduce the drop-out rate among the participants. As implied above, it is a frequent phenomenon with unsupervised tests to have high drop-out rates, i.e. participants who do not complete the test and thereby produce incomplete data sets. Incomplete data sets are not included in the analysis. We reduce the drop-out rate by placing a “high hurdle” (Reips, 2002a) at the beginning of the test in order to better control the self-selection of only sufficiently motivated participants. They have to go through introductory information about the test duration and tasks that await them, a demographic questionnaire, and a warm-up phase. In a warm-up phase, with example stimuli, participants get used to the instructions and can e.g. adjust the volume of their audio output device (headphones or loud speakers). During this phase, no suggested answers are given. The participants who follow through with the whole warm-up procedure are likely to also complete the main part of the test, the actual collection of ratings. The ones who are likely to drop out will probably do so during these early stages.

We emphasized in the instructions of all our tests that it is not the people that are being tested but the synthesis system, or if the synthetic nature of the data was not revealed, we emphasized that there is no correct or incorrect answer and that we were only interested in their spontaneous feelings toward the samples. Optimally, these reminders are repeated at various points of the test such as the introduction, the warm-up phase and before the main test. This should preclude that the participants feel overburdened and therefore drop out of the test.

The fourth aspect concerns thorough and careful pilot testing, which is a very important step when preparing a web-based test. The first thing to test is whether the technical components work reliably and do what they should be doing, the second thing is to ensure that the directions for the tasks are clearly understandable and that the work flow of listening and subsequent input of answers is ergonomic. It can be a good idea to provide a list of frequently asked questions to clarify known issues. The importance of serious pilot testing becomes clear when we imagine the typical situation a participant is in: Alone in front of their computer with no lab assistant around to ask just one quick question. So if anything is phrased ambiguously, or audio playback etc. does not work, the participant will be completely lost because it is very rare that they would write an email to the authors to ask for clarification.

If a test has a severe problem, it will become obvious since no one will be able to take the complete test. But there might also be hidden technical problems, misunderstandings or just a tiring way of presenting the speech samples which may confound the collected data. Thus, in pilot phases, we usually enhanced the tests by additional feedback text boxes at every single stimulus presentation, providing extensive opportunities for direct feedback via an uncomplicated channel. The demands for changes in phrasing of instructions, work flow, or the graphical layout were implemented (if possible) according to the feedback of the pilot testers before the big roll-out of the experiment. A smaller set of feedback opportunities was also implemented in the final version of a test. So, directly after the participants have completed the actual listening task, they can give general feedback by filling out text boxes with free text. We also collect basic technical feedback by asking the participants forced choice questions, e.g. regarding problems with audio play back.

B.3 Issues in data analysis and reporting of results

Important aspects when analyzing and reporting the data include measures for maintaining high data quality, reporting of drop-out rates and technical implementation details.

Quality assurance during data analysis can be done e.g. by using reaction times or rating means and ranges as outlier criteria. For instance, too long reaction times usually indicate that a participant has been doing something else besides listening to a sample and answering the associated questions. Similarly, if a participant's answers sum up to an unusually high mean of ratings, this might indicate that they did not seriously answer the questions, or they might have misunderstood the task. If this happens too often, however, there might be a problem with the test itself.

While these outlier treatments can also be found in non-web-based testing, the reporting of drop-out rates is a typical feature of internet-based tests that enhances the quality of data reporting. Where the logs permit a reliable counting of drop-outs, it is advisable to report the numbers. Other technical details which are useful to report include the manners of acquisition of participants, where and how long the test was online, and the software used for the implementation of the test.

B.4 Technical implementation

The technical implementation of web-based experiments does not require thorough programming skills since there is a variety of software, toolkits or web-services available. Typical technical examples nowadays include web-forms e.g. with an embedded audio player, or applications embedded into the web browser such as Java applets. The main task that remains to be done by the researcher is to adapt configuration files to the desired design of a given test. It is recommended as ‘good practice’ to use, if possible, existing test software since it helps avoiding basic mistakes regarding security issues. It has undergone software testing and complies with security demands, such as not disclosing personal data or the internal structure of the experiment through obvious naming of URLs (Reips, 2002a).

We inspected some free on-line toolkits or services and found that they differ in features such as capabilities of randomization and balancing of the test items, implementing certain sequence restrictions, playback of audio, recording of reaction times, visual design of the stimulus presentation slides, or the possibility to take the test in more than one session (flexible sign-in/sign-out). The latter may be helpful during longer test scenarios or when data shall be collected over a longer period of time. We finally decided to use WebExp 2.0 from CSTR (Keller et al., 2009) and PERCY from the Bavarian Archive for Speech Signals BAS (cf. e.g. Draxler, 2011), as described in Section 4.2.7.

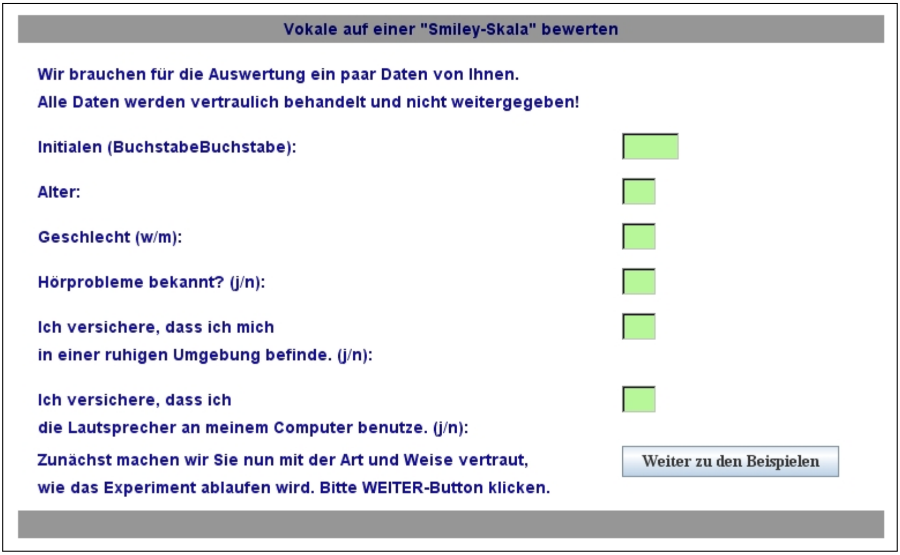
B.5 Screenshots of the web-based listening tests

The following figures are screenshots from the different listening tests. Explanations and translations of relevant slide items and text sections are provided along with the screenshots. The software package WebExp 2.0 (Keller et al., 2009) was used for the experiments in the Chapters 6 and 7, the software service PERCY (Draxler, 2011[1]) was used in the experiment in Chapter 11.

B.5.1 WebExp 2.0

Figures B.1 and B.2 show details of the perception study in the smiled vowel experiment, presented in Chapter 6.

[1] http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html



Vokale auf einer "Smiley-Skala" bewerten

Wir brauchen für die Auswertung ein paar Daten von Ihnen.
Alle Daten werden vertraulich behandelt und nicht weitergegeben!

Initialen (BuchstabeBuchstabe):

Alter:

Geschlecht (w/m):

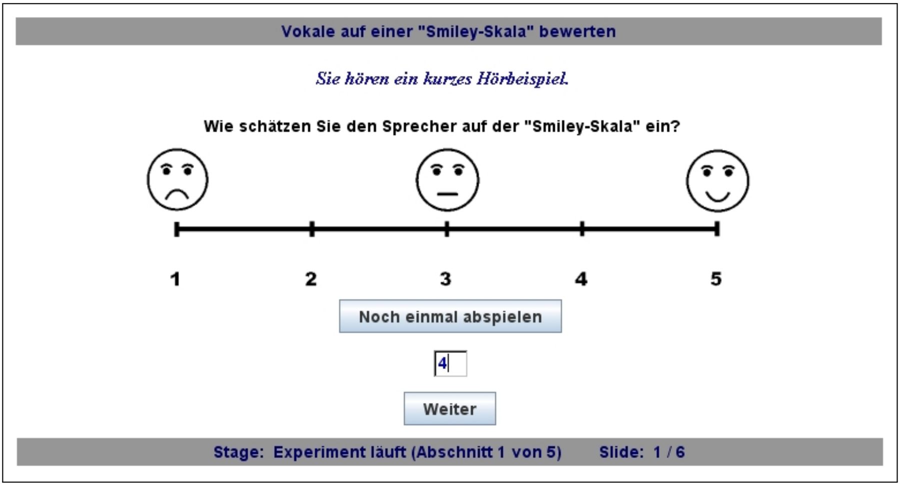
Hörprobleme bekannt? (j/n):

Ich versichere, dass ich mich
in einer ruhigen Umgebung befinde. (j/n):

Ich versichere, dass ich
die Lautsprecher an meinem Computer benutze. (j/n):

Zunächst machen wir Sie nun mit der Art und Weise vertraut,
wie das Experiment ablaufen wird. Bitte WEITER-Button klicken.

Figure B.1: Screenshot of the initial demographic questionnaire in the experiment on smiled vowels. Participants are asked to provide information about their name initials, age, gender, whether difficulty with hearing exists, a statement whether they are situated in a quiet atmosphere, and whether they are using the loudspeakers of their computers. The button on the lower right advances the slide to start with a warm-up phase (not depicted).



Vokale auf einer "Smiley-Skala" bewerten

Sie hören ein kurzes Hörbeispiel.

Wie schätzen Sie den Sprecher auf der "Smiley-Skala" ein?

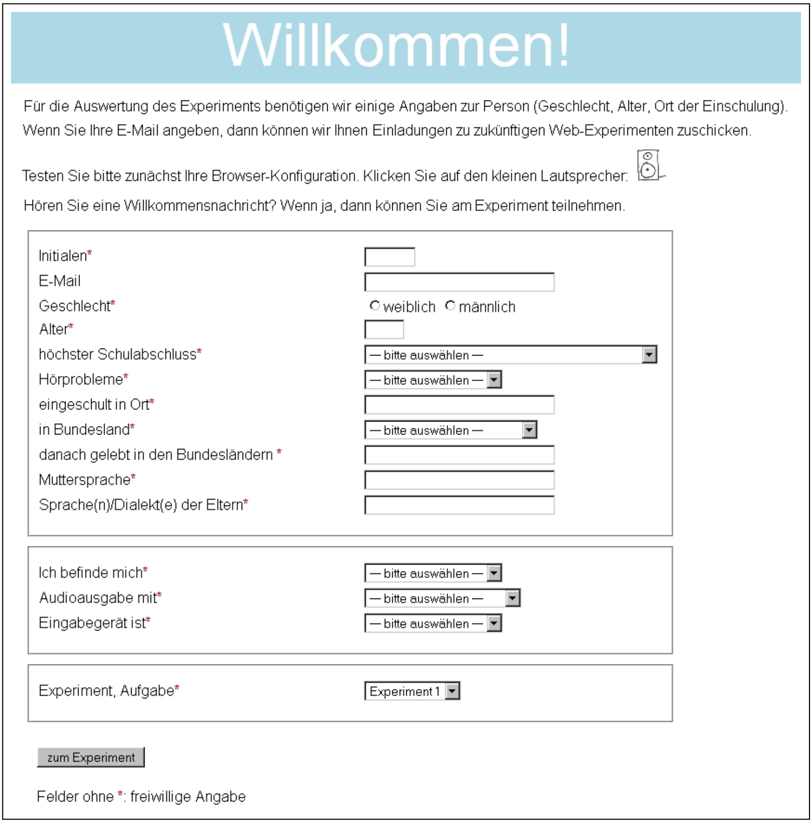
4

Stage: Experiment läuft (Abschnitt 1 von 5) Slide: 1 / 6

Figure B.2: Screenshot of a slide in the main phase of the experiment on smiled vowels. The participants have the opportunity to listen to a stimulus several times by pressing the button below the rating scale labeled "play again" („Noch einmal abspielen“). Then they have to enter a value between 1 and 5 into the small box above the 'next' button.

B.5.2 PERCY

Figures B.3 through B.6 depict details of the perception study in the regional accent experiment, presented in Chapter 11.



Willkommen!

Für die Auswertung des Experiments benötigen wir einige Angaben zur Person (Geschlecht, Alter, Ort der Einschulung). Wenn Sie Ihre E-Mail angeben, dann können wir Ihnen Einladungen zu zukünftigen Web-Experimenten zuschicken.

Testen Sie bitte zunächst Ihre Browser-Konfiguration. Klicken Sie auf den kleinen Lautsprecher:

Hören Sie eine Willkommensnachricht? Wenn ja, dann können Sie am Experiment teilnehmen.

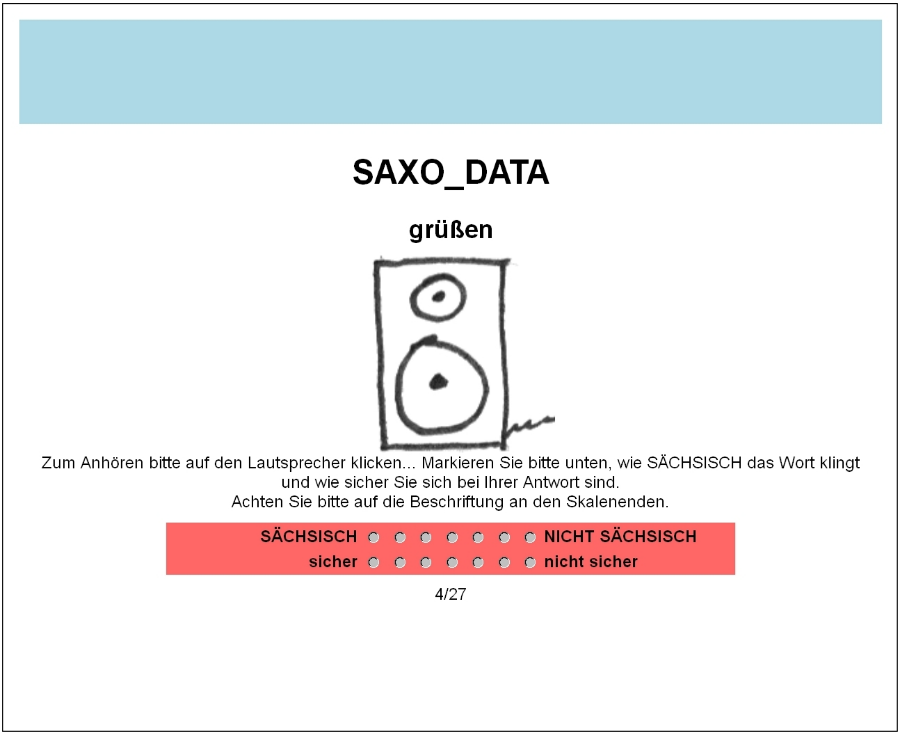
Initialen*	
E-Mail	
Geschlecht*	○ weiblich ○ männlich
Alter*	
höchster Schulabschluss*	— bitte auswählen —
Hörprobleme*	— bitte auswählen —
eingeschult in Ort*	
in Bundesland*	— bitte auswählen —
danach gelebt in den Bundesländern *	
Muttersprache*	
Sprache(n)/Dialekt(e) der Eltern*	

Ich befinde mich*	— bitte auswählen —
Audioausgabe mit*	— bitte auswählen —
Eingabegerät ist*	— bitte auswählen —

Experiment, Aufgabe*	Experiment 1

Felder ohne *: freiwillige Angabe

Figure B.3: Screenshot of initial instructions and demographic questionnaire for the regional accent experiment. Firstly, the participant is asked to test their internet browser by clicking on a small loudspeaker picture. If they hear a short message, their computer is correctly configured so that they can take part in the experiment. In the subsequent questionnaire, the participant is asked to provide information about their initials, gender, age, highest degree of education, and difficulty in hearing. Geographical information is asked to assess their regional background: Town and state of primary school, other federal states lived in, native language, and native language or dialects of parents. Furthermore, information about the test environment is demanded: Current surroundings (such as quiet office or internet cafe), kind of audio output (such as via loudspeakers or headphones), and input device (such as personal computer or hand-held device). Optionally, the participant may provide an email address if they are interested in receiving invitations for future experiments. Then, the participant has to select an experiment from a drop down menu according to the invitation they received by email. Finally, the gray button starts the experiment.



SAXO_DATA

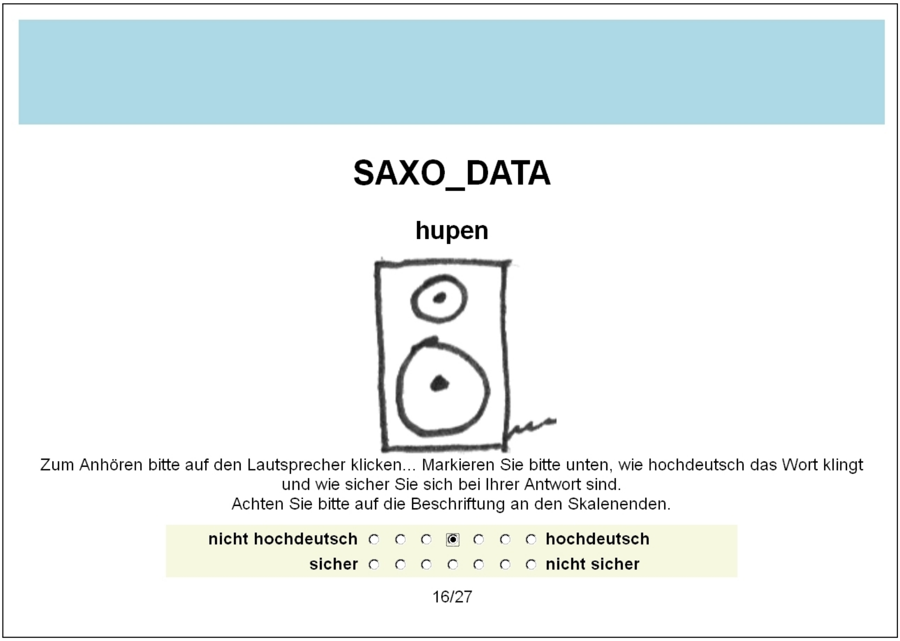
grüßen

Zum Anhören bitte auf den Lautsprecher klicken... Markieren Sie bitte unten, wie SÄCHSISCH das Wort klingt und wie sicher Sie sich bei Ihrer Antwort sind.
Achten Sie bitte auf die Beschriftung an den Skalenenden.

SÄCHSISCH ○ ○ ○ ○ ○ ○ ○ NICHT SÄCHSISCH
sicher ○ ○ ○ ○ ○ ○ ○ nicht sicher

4/27

Figure B.4: Screenshot of a typical rating task display. The wording of the stimulus is given above the drawing of the loudspeaker. The instructions are placed below the drawing, including that it is necessary to click on the loudspeaker for listening to the audio, and to pay attention to the labels at the end of the scales when providing the answers. The rating scales are inactive (gray filling in the circles) because the participant has not yet clicked on the loudspeaker picture to listen to the word. At the bottom, an indication is given about the progress of the experiment (slide 4 out of 27).



SAXO_DATA

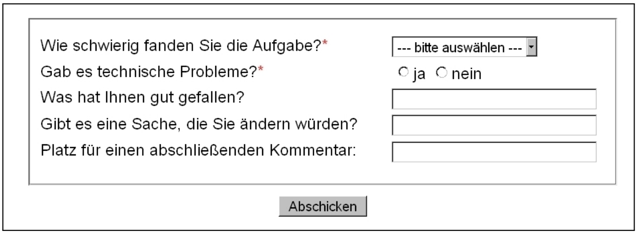
hupen

Zum Anhören bitte auf den Lautsprecher klicken... Markieren Sie bitte unten, wie hochdeutsch das Wort klingt und wie sicher Sie sich bei Ihrer Antwort sind.
Achten Sie bitte auf die Beschriftung an den Skalenenden.

nicht hochdeutsch	○	○	○	●	○	○	○	hochdeutsch
sicher	○	○	○	○	○	○	○	nicht sicher

16/27

Figure B.5: Screenshot of a typical rating task display. In contrast to Figure B.4, rating scales are now active, after having clicked on the loudspeaker picture to listen to the word, and the accent rating has been entered (black dot). A different background color is used to highlight that a different question is used than before, which is indicated by a different labeling at the end of the scales – we selected bright colors to ensure that participants realize the change of question type after the first half of the experiment.



Wie schwierig fanden Sie die Aufgabe?*	--- bitte auswählen ---
Gab es technische Probleme?*	○ ja ○ nein
Was hat Ihnen gut gefallen?	
Gibt es eine Sache, die Sie ändern würden?	
Platz für einen abschließenden Kommentar:	

Figure B.6: Closing questionnaire to collect feedback in general. The participant is asked to indicate how difficult the task was overall and whether they encountered technical problems. The last three text boxes allow for optional free-text feedback regarding positive and negative points about the whole experiment and a concluding general comment. Only few participants used these boxes.

Appendix C

Voice quality measurements

The following three figures present additional voice quality measurements of the synthetic speech tokens presented in Experiment I – Larynx height and voice quality (Chapter 5, p. 75).

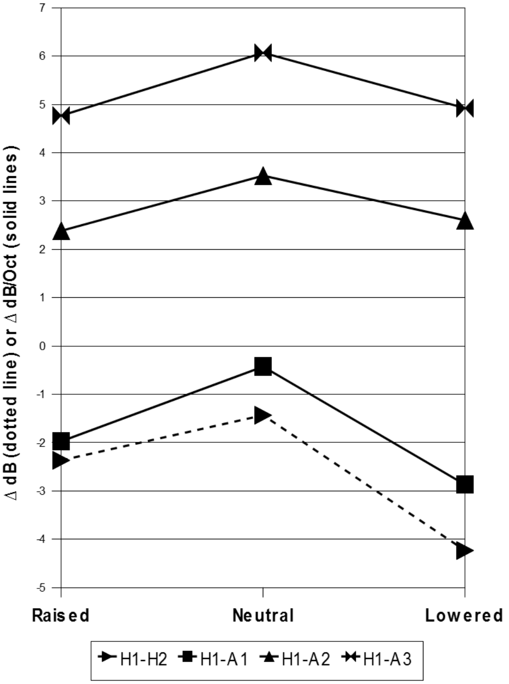


Figure C.1: Voice quality measurements when degree of breathiness is kept constant and larynx height is varied from lowered to raised setting: No breathiness added (amplitude of glottal gesture set to 0).

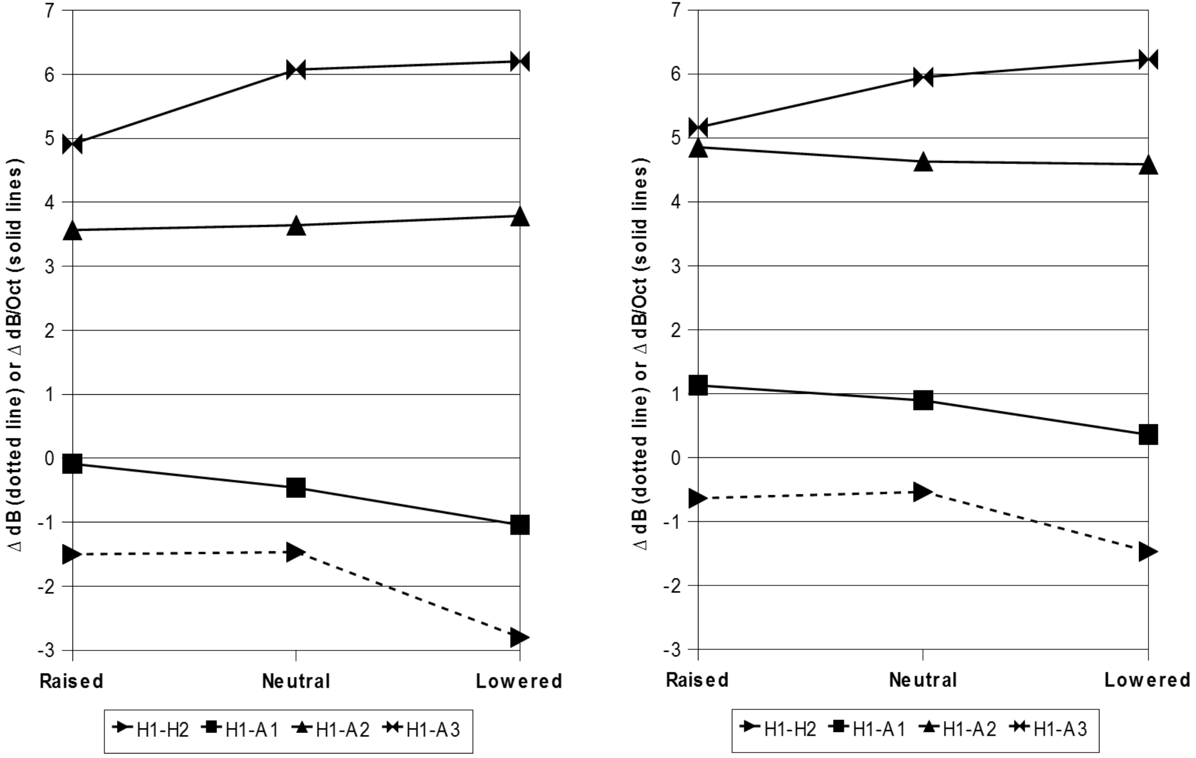


Figure C.2: **Left:** Slight breathiness added (amplitude of glottal gesture set to 5). **Right:** Moderate breathiness added (amplitude of glottal gesture set to 10).

Appendix D

Statistics of smiled vowel detection

The numbers given in Tables D.1 to D.3 were calculated using ANOVAs (SPSS: Unianova, $\alpha = 5\%$, see Chapter 6, p. 85). Dependent variable: Average rating per participant, averaged from 3 single ratings, based on 36 participants. Independent variables: vowel, lips, larynx, f_0. Significant values of p are printed in bold face.

Table D.1: Overall significance test.

Factors	df	F	p	Factors	df	F	p
vowel	3	122.381	**0.000**	vowel * lips * larynx	3	0.136	0.939
lips	1	4.560	**0.033**	vowel * f0	3	0.607	0.611
larynx	1	20.442	**0.000**	lips * f0	1	0.020	0.888
f0	1	73.871	**0.000**	vowel * lips * f0	3	0.658	0.578
vowel * lips	3	7.780	**0.000**	larynx * f0	1	0.170	0.680
vowel * larynx	3	2.232	0.083	vowel * larynx * f0	3	1.061	0.365
lips * larynx	1	2.031	0.154	lips * larynx * f0	1	0.002	0.965
				vowel * lips * larynx * f0	3	0.014	0.998

Table D.2: Post-hoc test (Scheffé) of the overall significance test.

Vowel	Vowel	p	Vowel	Vowel	p
/aː/	/iː/	**0.031**	/uː/	/aː/	**0.000**
	/uː/	**0.000**		/iː/	**0.000**
	/yː/	**0.000**		/yː/	0.808
Vowel	Vowel	p	Vowel	Vowel	p
/iː/	/aː/	**0.031**	/yː/	/aː/	**0.000**
	/uː/	**0.000**		/iː/	**0.000**
	/yː/	**0.000**		/uː/	0.808

Table D.3: Overall significance tests for each vowel.

	/aː/			/iː/		
Factors	df	F	p	df	F	p
lips	1	8.068	**0.005**	1	0.176	0.675
larynx	1	20.290	**0.000**	1	8.291	**0.004**
f0	1	14.126	**0.000**	1	17.818	**0.000**
lips * larynx	1	0.297	0.586	1	0.570	0.451
lips * f0	1	0.100	0.753	1	0.345	0.558
larynx * f0	1	0.185	0.667	1	0.007	0.933
lips * larynx * f0	1	0.040	0.841	1	0.001	0.978

	/uː/			/yː/		
Factors	df	F	p	df	F	p
lips	1	5.847	**0.016**	1	12.475	**0.000**
larynx	1	2.163	0.142	1	0.360	0.549
f0	1	24.608	**0.000**	1	17.541	**0.000**
lips * larynx	1	0.140	0.709	1	1.323	0.251
lips * f0	1	1.373	0.242	1	0.011	0.916
larynx * f0	1	1.880	0.171	1	1.026	0.312
lips * larynx * f0	1	0.006	0.940	1	0.002	0.968

Appendix E

Voice reports and statistics on aged vowel classification

The following supplementary material is referenced to at different points in Chapter 7. Tables E.1 to E.6 show the voice reports of the diphthongs in the age classification experiment before and after the telephone filter has been applied. Tables E.7 to E.12 show the acoustic properties of the basic vowels, integrated into the stimulus diphthongs, in the different voice profiles. Figure E.2 and Table E.14 depict the confusion of listeners between the age classes. Figure E.1 and Table E.13 show the number of listener judgments for each voice. Finally, Figure E.3 and Table E.15 show the classification behavior of an automatic age classifier, discussed in Section 7.4.3.

Table E.1: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([aɪ̯ aʊ̯ ɔɪ̯]). This table presents the age class **YOUNG**, *before* the telephone filter is applied. See also Section 7.3.2. Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	130.966	1.411	128.055	133.065	0.17	0.08	4.01	1.68	2.98	26.624	0
			aU	130.848	1.391	128.003	133.001	0.17	0.08	3.89	1.70	3.06	28.213	0
			OI	131.034	1.349	128.237	133.203	0.18	0.08	3.95	1.66	2.90	26.379	0
		Breathy	aI	130.942	1.424	128.071	133.051	0.19	0.09	4.14	1.80	3.19	27.559	0
			aU	130.87	1.393	128.008	133.071	0.18	0.09	4.22	1.84	3.31	28.443	0
			OI	131.004	1.363	128.249	133.187	0.20	0.10	4.07	1.76	3.10	27.288	0
	Irregular	Modal	aI	130.964	1.447	127.818	133.276	0.22	0.10	4.03	1.74	3.03	25.863	0
			aU	130.846	1.427	127.791	133.212	0.22	0.10	3.91	1.72	3.09	27.263	0
			OI	131.032	1.385	128.001	133.416	0.23	0.11	4.01	1.72	2.94	25.645	0
		Breathy	aI	130.94	1.46	127.84	133.246	0.24	0.11	4.15	1.83	3.23	26.84	0
			aU	130.868	1.429	127.751	133.284	0.24	0.11	4.23	1.85	3.32	27.528	0
			OI	131.002	1.399	128.019	133.397	0.25	0.12	4.10	1.78	3.12	26.584	0
Mid	Regular	Modal	aI	120.955	1.412	118.082	123.157	0.20	0.09	4.10	1.79	2.96	25.554	0
			aU	120.841	1.392	117.96	122.924	0.20	0.10	4.05	1.86	3.13	27.244	0
			OI	121.024	1.36	118.232	123.169	0.19	0.10	4.00	1.76	2.96	25.987	0
		Breathy	aI	120.944	1.41	118.029	123.156	0.22	0.10	4.20	1.90	3.20	26.644	0
			aU	120.865	1.389	117.956	123.025	0.20	0.10	4.38	2.03	3.41	27.715	0
			OI	120.991	1.384	118.208	123.228	0.21	0.10	4.02	1.82	3.09	27.165	0
	Irregular	Modal	aI	120.953	1.447	117.801	123.36	0.26	0.12	4.20	1.83	3.02	24.742	0
			aU	120.839	1.428	117.701	123.132	0.26	0.13	4.10	1.88	3.15	26.29	0
			OI	121.021	1.397	117.989	123.376	0.25	0.12	4.10	1.80	3.01	25.268	0
		Breathy	aI	120.942	1.446	117.751	123.361	0.27	0.13	4.26	1.92	3.24	25.879	0
			aU	120.863	1.425	117.697	123.234	0.26	0.13	4.40	2.04	3.41	26.841	0
			OI	120.989	1.42	117.963	123.436	0.26	0.13	4.07	1.85	3.12	26.409	0
Low	Regular	Modal	aI	110.936	1.426	108.035	113.256	0.21	0.09	4.21	1.92	3.09	25.001	0
			aU	110.833	1.398	107.975	113.084	0.22	0.10	4.24	2.06	3.29	26.662	0
			OI	111.009	1.364	108.208	113.264	0.22	0.10	4.45	2.08	3.23	24.677	0
		Breathy	aI	110.934	1.399	108.005	113.23	0.23	0.11	4.35	2.03	3.23	26.192	0
			aU	110.859	1.393	107.953	113.081	0.23	0.11	4.54	2.24	3.57	27.314	0
			OI	110.981	1.369	108.164	113.252	0.23	0.11	4.51	2.13	3.33	26.147	0
	Irregular	Modal	aI	110.934	1.462	107.773	113.474	0.28	0.12	4.29	1.96	3.14	24.196	0
			aU	110.831	1.434	107.726	113.293	0.28	0.13	4.29	2.07	3.32	25.69	0
			OI	111.006	1.4	107.97	113.494	0.28	0.13	4.56	2.14	3.27	23.959	0
		Breathy	aI	110.932	1.434	107.756	113.456	0.30	0.14	4.39	2.05	3.23	25.457	0
			aU	110.857	1.43	107.709	113.301	0.30	0.14	4.56	2.26	3.58	26.415	0
			OI	110.978	1.404	107.925	113.478	0.29	0.14	4.54	2.16	3.34	25.44	0

Table E.2: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([aɪ̯ aʊ̯ ɔɪ̯]). This table presents the age class **ADULT**, *before* the telephone filter is applied. See also Section 7.3.2. Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	115.893	1.481	112.515	118.826	0.37	0.19	5.14	2.48	3.99	25.07	0
			aU	115.808	1.477	112.524	118.482	0.37	0.19	5.48	2.66	4.27	25.917	0
			OI	115.909	1.472	112.576	118.602	0.35	0.18	5.29	2.55	4.11	25.103	0
		Breathy	aI	115.891	1.489	112.536	118.849	0.38	0.20	5.53	2.70	4.35	25.175	0
			aU	115.811	1.478	112.446	118.497	0.37	0.19	5.61	2.76	4.40	25.498	0
			OI	115.912	1.464	112.55	118.594	0.37	0.19	5.65	2.73	4.39	25.119	0
	Irregular	Modal	aI	115.891	1.531	112.28	119.022	0.43	0.22	5.17	2.49	4.00	24.442	0
			aU	115.806	1.528	112.282	118.703	0.44	0.23	5.50	2.68	4.28	25.229	0
			OI	115.907	1.521	112.343	118.823	0.42	0.21	5.31	2.56	4.11	24.529	0
		Breathy	aI	115.889	1.54	112.304	119.061	0.44	0.23	5.56	2.71	4.36	24.595	0
			aU	115.809	1.529	112.188	118.707	0.44	0.23	5.61	2.77	4.40	24.791	0
			OI	115.91	1.514	112.317	118.818	0.43	0.23	5.67	2.74	4.40	24.538	0
Mid	Regular	Modal	aI	108.385	1.481	104.911	111.356	0.38	0.19	5.45	2.81	4.26	24.686	0
			aU	108.298	1.478	104.901	110.946	0.39	0.19	5.67	2.96	4.48	25.584	0
			OI	108.401	1.464	105.007	111.01	0.39	0.20	5.56	2.90	4.33	24.608	0
		Breathy	aI	108.384	1.502	104.894	111.337	0.40	0.20	5.82	3.03	4.59	24.836	0
			aU	108.299	1.476	104.91	110.937	0.39	0.20	5.89	3.07	4.65	25.19	0
			OI	108.406	1.473	104.958	111.041	0.40	0.21	5.85	3.05	4.57	24.691	0
	Irregular	Modal	aI	108.383	1.532	104.651	111.562	0.45	0.22	5.48	2.82	4.27	24.06	0
			aU	108.295	1.529	104.644	111.166	0.46	0.23	5.71	2.97	4.50	24.884	0
			OI	108.399	1.514	104.74	111.234	0.45	0.23	5.57	2.91	4.34	24.001	0
		Breathy	aI	108.382	1.553	104.628	111.547	0.47	0.23	5.84	3.04	4.60	24.249	0
			aU	108.297	1.528	104.651	111.159	0.47	0.24	5.91	3.08	4.66	24.467	0
			OI	108.404	1.524	104.687	111.256	0.47	0.25	5.86	3.06	4.58	24.1	0
Low	Regular	Modal	aI	100.885	1.484	97.502	104.074	0.43	0.19	5.88	3.12	4.27	24.244	0
			aU	100.789	1.48	97.462	103.496	0.41	0.19	6.01	3.25	4.48	25.238	0
			OI	100.899	1.465	97.556	103.425	0.40	0.18	5.95	3.16	4.33	24.174	0
		Breathy	aI	100.886	1.514	97.45	103.898	0.43	0.19	6.10	3.28	4.48	24.51	0
			aU	100.787	1.475	97.448	103.561	0.41	0.18	6.33	3.43	4.70	25.045	0
			OI	100.897	1.484	97.562	103.526	0.41	0.19	6.11	3.31	4.52	24.445	0
	Irregular	Modal	aI	100.883	1.534	97.254	104.282	0.50	0.22	5.93	3.15	4.30	23.584	0
			aU	100.787	1.53	97.221	103.725	0.49	0.22	6.03	3.27	4.49	24.508	0
			OI	100.897	1.515	97.321	103.644	0.47	0.21	6.00	3.19	4.36	23.553	0
		Breathy	aI	100.884	1.563	97.205	104.106	0.50	0.23	6.12	3.31	4.50	23.9	0
			aU	100.785	1.525	97.21	103.794	0.48	0.21	6.38	3.45	4.72	24.338	0
			OI	100.895	1.534	97.315	103.738	0.49	0.22	6.14	3.32	4.53	23.869	0

Table E.3: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([a͜ɪ a͜ʊ ɔ͜ɪ]). This table presents the age class **Senior**, *before* the telephone filter is applied. See also Section 7.3.2. Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	150.898	2.054	145.436	155.531	0.63	0.29	5.65	2.34	4.63	23.15	0
			aU	150.81	2.047	145.323	155.481	0.60	0.28	5.80	2.43	4.73	23.435	0
			OI	150.915	2.046	145.568	155.534	0.64	0.30	5.77	2.37	4.68	23.189	0
		Breathy	aI	150.874	2.065	145.454	155.421	0.57	0.25	5.89	2.46	4.78	22.215	0
			aU	150.786	2.042	145.335	155.417	0.56	0.25	6.11	2.57	4.96	22.701	0
			OI	150.88	2.069	145.486	155.457	0.57	0.26	5.80	2.42	4.68	22.618	0
	Irregular	Modal	aI	150.893	2.323	144.683	156.422	0.77	0.35	5.73	2.36	4.66	22.054	0
			aU	150.801	2.311	144.547	156.322	0.74	0.34	5.86	2.46	4.77	22.318	0
			OI	150.909	2.313	144.792	156.399	0.79	0.37	5.86	2.39	4.71	22.139	0
		Breathy	aI	150.865	2.332	144.697	156.298	0.70	0.31	6.01	2.51	4.84	21.028	0
			aU	150.784	2.312	144.551	156.303	0.69	0.31	6.21	2.61	5.00	21.499	0
			OI	150.871	2.332	144.732	156.343	0.71	0.32	5.90	2.46	4.73	21.445	0
Mid	Regular	Modal	aI	128.369	2.042	122.988	132.888	0.78	0.38	6.27	3.15	5.43	21.601	0
			aU	128.265	2.057	122.96	133.035	0.83	0.41	6.26	3.03	5.31	21.496	0
			OI	128.378	2.038	123.086	132.9	0.76	0.37	6.29	3.10	5.38	21.769	0
		Breathy	aI	128.365	2.035	122.911	132.924	0.73	0.34	6.59	3.23	5.54	20.401	0
			aU	128.25	2.034	122.934	132.949	0.72	0.32	6.95	3.37	5.90	20.949	0
			OI	128.365	2.032	122.959	132.913	0.73	0.35	6.79	3.34	5.72	20.951	0
	Irregular	Modal	aI	128.365	2.308	122.19	133.766	0.96	0.46	6.30	3.21	5.50	20.479	0
			aU	128.259	2.327	122.217	133.968	1.01	0.50	6.28	3.04	5.33	20.333	0
			OI	128.371	2.304	122.344	133.737	0.93	0.45	6.34	3.14	5.44	20.656	0
		Breathy	aI	128.361	2.304	122.168	133.81	0.90	0.42	6.70	3.32	5.62	19.174	0
			aU	128.247	2.306	122.227	133.862	0.88	0.39	6.96	3.38	5.91	19.724	0
			OI	128.362	2.3	122.216	133.766	0.89	0.43	6.94	3.45	5.80	19.732	0
Low	Regular	Modal	aI	105.845	2.057	100.195	110.209	1.01	0.51	7.16	3.73	5.30	20.595	0
			aU	105.743	2.031	100.164	110.218	1.04	0.53	7.19	3.83	5.43	20.665	0
			OI	105.855	2.046	100.177	110.243	1.02	0.51	7.17	3.65	5.21	20.568	0
		Breathy	aI	105.839	2.063	100.346	110.263	0.96	0.46	8.01	4.17	5.95	18.559	0
			aU	105.7	2.043	100.392	110.215	0.91	0.40	7.78	4.17	5.97	19.316	0
			OI	105.832	2.079	100.466	110.315	0.98	0.48	7.98	4.06	5.80	19.008	0
	Irregular	Modal	aI	105.841	2.325	99.449	110.854	1.25	0.63	7.18	3.74	5.30	19.449	0
			aU	105.741	2.306	99.398	111.006	1.29	0.66	7.19	3.83	5.44	19.515	0
			OI	105.851	2.314	99.39	110.869	1.26	0.64	7.17	3.65	5.18	19.398	0
		Breathy	aI	105.836	2.331	99.578	110.944	1.20	0.59	8.08	4.19	5.95	17.328	0
			aU	105.696	2.313	99.646	111.05	1.14	0.51	7.79	4.17	5.96	18.092	0
			OI	105.824	2.344	99.738	111.015	1.22	0.60	8.03	4.07	5.80	17.781	0

Table E.4: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([aɪ̯ aʊ̯ ɔɪ̯]). This table presents the age class **Young**, *after* the telephone filter is applied. See also Section 7.3.2. Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	130.962	1.411	128.054	133.072	0.182	0.085	4.550	1.934	3.368	25.89	0
			aU	130.841	1.388	128.001	132.997	0.184	0.092	4.360	1.885	3.375	27.632	0
			OI	131.032	1.347	128.23	133.2	0.198	0.096	4.492	1.889	3.263	25.657	0
		Breathy	aI	130.934	1.423	128.075	133.042	0.199	0.095	4.970	2.167	3.817	26.628	0
			aU	130.861	1.389	128.006	133.061	0.191	0.094	4.943	2.139	3.855	27.472	0
			OI	130.998	1.36	128.25	133.181	0.218	0.110	4.777	2.070	3.603	26.394	0
	Irregular	Modal	aI	130.959	1.447	127.819	133.28	0.231	0.110	4.591	2.006	3.430	25.112	0
			aU	130.839	1.424	127.788	133.207	0.240	0.120	4.355	1.893	3.377	26.642	0
			OI	131.03	1.383	128.004	133.414	0.248	0.122	4.567	1.956	3.306	24.913	0
		Breathy	aI	130.932	1.458	127.843	133.25	0.251	0.120	4.986	2.209	3.859	25.881	0
			aU	130.859	1.426	127.752	133.275	0.248	0.123	4.970	2.151	3.866	26.498	0
			OI	130.996	1.396	128.023	133.393	0.269	0.138	4.813	2.097	3.636	25.676	0
Mid	Regular	Modal	aI	120.951	1.412	118.082	123.157	0.228	0.114	4.690	2.032	3.345	24.798	0
			aU	120.835	1.391	117.958	122.918	0.221	0.113	4.600	2.099	3.528	26.641	0
			OI	121.022	1.36	118.222	123.166	0.230	0.119	4.419	1.926	3.296	25.267	0
		Breathy	aI	120.935	1.409	118.025	123.15	0.233	0.114	5.047	2.276	3.843	25.513	0
			aU	120.855	1.387	117.96	123.022	0.212	0.109	5.134	2.401	4.021	26.648	0
			OI	120.985	1.382	118.187	123.225	0.225	0.114	4.586	2.095	3.550	26.276	0
	Irregular	Modal	aI	120.949	1.448	117.801	123.361	0.289	0.143	4.833	2.103	3.411	23.961	0
			aU	120.832	1.427	117.698	123.127	0.285	0.146	4.668	2.131	3.557	25.646	0
			OI	121.019	1.397	117.94	123.375	0.271	0.137	4.492	1.997	3.339	24.543	0
		Breathy	aI	120.933	1.445	117.751	123.356	0.294	0.144	5.141	2.330	3.891	24.72	0
			aU	120.853	1.423	117.705	123.231	0.274	0.141	5.140	2.398	4.002	25.715	0
			OI	120.982	1.419	117.917	123.436	0.285	0.143	4.654	2.123	3.595	25.507	0
Low	Regular	Modal	aI	110.931	1.425	108.038	113.26	0.234	0.112	5.005	2.250	3.616	24.246	0
			aU	110.827	1.396	107.975	113.084	0.223	0.100	4.798	2.311	3.689	26.05	0
			OI	111.005	1.363	108.207	113.266	0.252	0.128	5.260	2.392	3.672	23.909	0
		Breathy	aI	110.925	1.396	108.008	113.234	0.241	0.117	5.264	2.451	3.873	24.93	0
			aU	110.85	1.391	107.953	113.076	0.232	0.112	5.479	2.678	4.258	26.094	0
			OI	110.975	1.366	108.17	113.256	0.244	0.123	5.450	2.524	3.908	25.106	0
	Irregular	Modal	aI	110.928	1.46	107.774	113.474	0.304	0.146	5.099	2.327	3.669	23.43	0
			aU	110.824	1.433	107.726	113.298	0.290	0.130	4.847	2.338	3.731	25.039	0
			OI	111.002	1.399	107.946	113.495	0.319	0.162	5.365	2.463	3.690	23.172	0
		Breathy	aI	110.922	1.432	107.76	113.461	0.309	0.150	5.302	2.489	3.882	24.175	0
			aU	110.847	1.428	107.711	113.296	0.301	0.145	5.498	2.705	4.272	25.132	0
			OI	110.972	1.401	107.91	113.483	0.309	0.154	5.506	2.576	3.906	24.39	0

Table E.5: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([aɪ̯ aʊ̯ ɔɪ̯]). This table presents the age class **ADULT**, *after* the telephone filter is applied. See also Section 7.3.2. Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	115.881	1.477	112.516	118.819	0.384	0.208	6.686	3.265	5.196	23.267	0
			aU	115.795	1.474	112.527	118.504	0.369	0.193	7.001	3.457	5.519	24.046	0
			OI	115.9	1.466	112.583	118.627	0.373	0.199	6.712	3.265	5.215	23.67	0
		Breathy	aI	115.879	1.486	112.532	118.821	0.393	0.211	7.121	3.493	5.604	23.633	0
			aU	115.8	1.475	112.453	118.497	0.372	0.191	7.156	3.553	5.647	23.602	0
			OI	115.906	1.458	112.559	118.576	0.383	0.206	7.022	3.411	5.440	23.757	0
	Irregular	Modal	aI	115.879	1.527	112.285	119.004	0.447	0.240	6.750	3.299	5.224	22.603	0
			aU	115.793	1.525	112.282	118.727	0.437	0.229	7.035	3.487	5.545	23.277	0
			OI	115.897	1.516	112.352	118.854	0.435	0.229	6.744	3.285	5.221	23.058	0
		Breathy	aI	115.877	1.537	112.3	119.087	0.457	0.243	7.176	3.518	5.624	23.025	0
			aU	115.797	1.526	112.196	118.704	0.440	0.227	7.161	3.570	5.645	22.827	0
			OI	115.903	1.509	112.325	118.798	0.447	0.237	7.040	3.418	5.450	23.155	0
Mid	Regular	Modal	aI	108.375	1.479	104.854	111.417	0.396	0.203	7.133	3.598	5.452	22.695	0
			aU	108.285	1.477	104.903	110.938	0.382	0.193	7.464	3.803	5.720	23.466	0
			OI	108.392	1.458	104.949	111.018	0.397	0.207	7.233	3.720	5.551	22.89	0
		Breathy	aI	108.37	1.502	104.886	111.407	0.404	0.205	7.635	3.919	5.957	23.067	0
			aU	108.288	1.475	104.912	110.931	0.390	0.200	7.780	3.985	5.991	23.003	0
			OI	108.397	1.466	104.912	111.046	0.417	0.221	7.440	3.850	5.751	23.097	0
	Irregular	Modal	aI	108.373	1.53	104.588	111.619	0.465	0.238	7.179	3.625	5.482	22.012	0
			aU	108.283	1.528	104.647	111.159	0.455	0.229	7.525	3.843	5.757	22.678	0
			OI	108.39	1.508	104.674	111.238	0.468	0.244	7.274	3.737	5.550	22.246	0
		Breathy	aI	108.368	1.552	104.622	111.611	0.474	0.240	7.690	3.942	5.984	22.436	0
			aU	108.286	1.527	104.657	111.154	0.468	0.239	7.847	4.012	6.015	22.194	0
			OI	108.395	1.517	104.632	111.257	0.487	0.259	7.480	3.866	5.766	22.464	0
Low	Regular	Modal	aI	100.876	1.481	97.507	104.303	0.431	0.197	8.165	4.291	5.836	21.921	0
			aU	100.777	1.477	97.466	103.497	0.405	0.182	8.063	4.360	5.971	22.899	0
			OI	100.889	1.46	97.571	103.426	0.413	0.190	8.192	4.368	5.954	22.11	0
		Breathy	aI	100.873	1.513	97.45	104.07	0.433	0.202	8.524	4.573	6.222	22.263	0
			aU	100.774	1.472	97.451	103.577	0.403	0.179	8.763	4.733	6.427	22.603	0
			OI	100.885	1.478	97.562	103.521	0.421	0.198	8.311	4.505	6.084	22.598	0
	Irregular	Modal	aI	100.874	1.532	97.258	104.514	0.506	0.228	8.225	4.334	5.874	21.199	0
			aU	100.775	1.529	97.224	103.73	0.482	0.214	8.081	4.363	5.974	22.078	0
			OI	100.886	1.511	97.331	103.652	0.483	0.219	8.300	4.417	6.020	21.44	0
		Breathy	aI	100.871	1.563	97.206	104.273	0.509	0.234	8.564	4.609	6.259	21.606	0
			aU	100.772	1.523	97.217	103.813	0.479	0.211	8.864	4.784	6.472	21.805	0
			OI	100.883	1.528	97.316	103.74	0.495	0.229	8.355	4.531	6.109	21.98	0

Table E.6: Voice report details of the diphthongs used for stimulus creation. Each age profile is represented by the 3 diphthongs ([aɪ̯ aʊ̯ ɔɪ̯]). This table presents the age class **Senior**, *after* the telephone filter is applied. See also Section 7.3.2.Measured with Praat (Boersma, 2001), using standard settings.

f_0 level	Irregularity	Breathiness	Vowel	f_0 mean	f_0 SD	f_0 min	f_0 max	Jitter (local)	Jitter (rap)	Shimmer (local)	Shimmer (apq3)	Shimmer (apq5)	Mean HNR [Hz]	Frames unvoiced
High	Regular	Modal	aI	150.88	2.056	145.414	155.564	0.638	0.290	7.091	2.882	5.690	21.551	0
			aU	150.792	2.046	145.29	155.501	0.615	0.285	7.337	3.097	5.930	21.096	0
			OI	150.905	2.043	145.586	155.565	0.652	0.303	7.249	2.917	5.738	21.63	0
		Breathy	aI	150.86	2.062	145.476	155.408	0.573	0.252	6.890	2.858	5.496	21.213	0
			aU	150.776	2.039	145.343	155.409	0.558	0.251	7.147	3.004	5.711	21.274	0
			OI	150.871	2.066	145.503	155.447	0.577	0.259	6.696	2.761	5.278	21.764	0
	Irregular	Modal	aI	150.874	2.325	144.667	156.471	0.779	0.353	7.270	2.954	5.785	20.401	0
			aU	150.783	2.309	144.523	156.339	0.750	0.347	7.525	3.227	6.084	19.931	0
			OI	150.899	2.312	144.811	156.434	0.801	0.375	7.431	2.978	5.813	20.549	0
		Breathy	aI	150.852	2.328	144.727	156.281	0.701	0.308	7.089	2.975	5.607	20	0
			aU	150.774	2.31	144.56	156.301	0.684	0.309	7.346	3.122	5.812	20.014	0
			OI	150.862	2.33	144.751	156.336	0.709	0.318	6.868	2.836	5.366	20.569	0
Mid	Regular	Modal	aI	128.353	2.038	122.895	132.958	0.792	0.389	8.619	4.332	7.323	19.66	0
			aU	128.248	2.054	122.923	133.148	0.841	0.413	8.188	3.882	6.778	18.8	0
			OI	128.364	2.03	123.078	132.951	0.772	0.381	8.241	4.096	7.055	20.105	0
		Breathy	aI	128.358	2.037	122.922	132.957	0.729	0.335	8.145	3.968	6.684	19.24	0
			aU	128.239	2.037	122.804	132.938	0.723	0.321	8.568	4.107	7.162	19.375	0
			OI	128.358	2.031	122.961	132.881	0.720	0.341	8.336	4.078	6.835	20.119	0
	Irregular	Modal	aI	128.351	2.303	122.145	133.818	0.966	0.473	8.786	4.479	7.448	18.464	0
			aU	128.241	2.325	122.189	134.083	1.025	0.505	8.178	3.884	6.750	17.557	0
			OI	128.358	2.299	122.333	133.788	0.943	0.466	8.343	4.194	7.131	18.936	0
		Breathy	aI	128.353	2.307	121.996	133.806	0.893	0.416	8.352	4.133	6.796	17.976	0
			aU	128.235	2.307	121.981	133.825	0.882	0.392	8.627	4.129	7.179	18.062	0
			OI	128.353	2.3	122.206	133.72	0.885	0.421	8.591	4.279	6.965	18.867	0
Low	Regular	Modal	aI	105.83	2.059	100.276	110.196	1.010	0.507	9.967	5.183	7.315	17.779	0
			aU	105.728	2.035	100.234	110.204	1.031	0.525	10.144	5.526	7.752	17.181	0
			OI	105.843	2.044	100.209	110.227	1.023	0.519	9.612	5.026	7.040	18.11	0
		Breathy	aI	105.83	2.063	100.359	110.262	0.957	0.454	10.179	5.194	7.382	17.327	0
			aU	105.691	2.041	100.405	110.225	0.908	0.398	9.440	5.103	7.243	17.505	0
			OI	105.829	2.077	100.494	110.314	0.978	0.476	9.875	5.006	7.089	17.948	0
	Irregular	Modal	aI	105.825	2.328	99.539	110.825	1.248	0.632	10.089	5.235	7.337	16.529	0
			aU	105.727	2.31	99.462	110.995	1.287	0.659	10.070	5.448	7.586	15.882	0
			OI	105.838	2.316	99.438	110.845	1.265	0.642	9.730	5.035	7.016	16.835	0
		Breathy	aI	105.829	2.331	99.599	110.901	1.196	0.578	10.404	5.262	7.421	16	0
			aU	105.688	2.311	99.668	111.073	1.130	0.503	9.492	5.145	7.256	16.172	0
			OI	105.822	2.343	99.776	110.975	1.210	0.596	10.061	5.028	7.084	16.638	0

Table E.7: F_1, F_2, F_3, and f_0 in **vowel [a]** of [a͜ɪ] in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f_0 level	Irregularity	Breathiness	f_0 mean [Hz]	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
[a] in [a͜ɪ]	Young	High	Regular	Modal	128	711	1275	2698
				Breathy	128	757	1291	2693
			Irregular	m	128	711	1276	2698
				b	128	759	1293	2696
		Mid	r	m	119	687	1292	2706
				b	118	705	1291	2726
			i	m	118	689	1292	2703
				b	118	708	1289	2724
		Low	r	m	108	700	1278	2701
				b	108	748	1318	2718
			i	m	108	702	1280	2698
				b	108	750	1323	2716
	Adult	High	Regular	m	113	734	1296	2634
				b	113	623	1230	2640
			Irregular	m	113	734	1298	2645
				b	113	632	1232	2643
		Mid	r	m	105	705	1283	2641
				b	105	623	1204	2657
			i	m	105	699	1282	2637
				b	105	618	1201	2651
		Low	r	m	98	674	1278	2632
				b	98	711	1162	2620
			i	m	98	668	1276	2632
				b	98	713	1161	2618
	Senior	High	Regular	m	147	591	1239	2618
				b	147	562	1281	2602
			Irregular	m	146	600	1244	2614
				b	146	558	1281	2599
		Mid	r	m	124	582	1231	2618
				b	124	561	1235	2607
			i	m	124	550	1223	2610
				b	124	566	1234	2604
		Low	r	m	102	694	1184	2585
				b	102	497	1234	2607
			i	m	101	686	1182	2582
				b	101	497	1237	2600

Table E.8: F_1, F_2, F_3, and f_0 in **vowel [a]** of [a͜ʊ] in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f_0 level	Irregularity	Breathiness	f_0 mean [Hz]	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
[a] in [a͜ʊ]	Young	High	Regular	Modal	128	704	1243	2732
				Breathy	128	742	1264	2745
			Irregular	m	128	704	1242	2738
				b	128	743	1263	2750
		Mid	r	m	118	689	1276	2724
				b	118	721	1278	2741
			i	m	118	690	1276	2722
				b	118	718	1277	2746
		Low	r	m	108	689	1232	2715
				b	108	709	1230	2741
			i	m	108	689	1231	2716
				b	108	709	1232	2745
	Adult	High	Regular	m	113	716	1243	2689
				b	113	629	1207	2686
			Irregular	m	113	716	1243	2696
				b	113	627	1206	2688
		Mid	r	m	105	686	1240	2645
				b	105	637	1191	2671
			i	m	105	684	1238	2643
				b	105	635	1191	2670
		Low	r	m	98	671	1253	2650
				b	98	650	1124	2639
			i	m	98	669	1252	2648
				b	98	652	1124	2636
	Senior	High	Regular	m	147	531	1180	2636
				b	147	529	1206	2625
			Irregular	m	146	542	1182	2627
				b	146	528	1210	2618
		Mid	r	m	124	618	1228	2636
				b	124	542	1224	2646
			i	m	123	606	1224	2639
				b	123	536	1219	2649
		Low	r	m	101	629	1150	2613
				b	101	476	1211	2625
			i	m	101	585	1135	2601
				b	101	474	1206	2616

Table E.9: F_1, F_2, F_3, and f_0 in **vowel [ɪ]** of [a͜ɪ] in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f_0 level	Irregularity	Breathiness	f_0 mean [Hz]	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
[ɪ] in [a͜ɪ]	YOUNG	High	Regular	Modal	131	384	2077	2906
				Breathy	131	391	2088	2903
			Irregular	m	131	385	2078	2906
				b	131	391	2089	2903
		Mid	r	m	121	389	2064	2918
				b	121	376	2089	3003
			i	m	121	389	2064	2915
				b	121	376	2088	2999
		Low	r	m	111	407	2067	2905
				b	111	398	2111	2903
			i	m	111	407	2068	2904
				b	111	398	2112	2905
	ADULT	High	Regular	m	116	415	2096	2927
				b	116	399	2084	2952
			Irregular	m	116	415	2097	2930
				b	116	399	2084	2951
		Mid	r	m	109	411	2063	2905
				b	109	404	2052	2915
			i	m	109	410	2064	2905
				b	109	404	2052	2915
		Low	r	m	101	391	2048	2979
				b	101	390	2059	2931
			i	m	101	390	2049	2983
				b	101	391	2058	2930
	SENIOR	High	Regular	m	151	403	2054	2901
				b	151	410	2089	2894
			Irregular	m	151	403	2057	2899
				b	151	410	2089	2890
		Mid	r	m	129	382	2056	2882
				b	129	405	2052	2876
			i	m	129	382	2058	2887
				b	129	406	2052	2878
		Low	r	m	106	393	1997	2835
				b	106	413	2038	2869
			i	m	106	393	2000	2834
				b	106	414	2041	2870

Table E.10: F_1, F_2, F_3, and f_0 in **vowel [ɪ]** of [ɔ͜ɪ] in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f_0 level	Irregularity	Breathiness	f_0 mean [Hz]	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
[ɪ] in [ɔ͜ɪ]	Young	High	Regular	Modal	131	376	2094	2964
				Breathy	130	384	2093	2972
			Irregular	m	131	377	2094	2963
				b	130	384	2094	2975
		Mid	r	m	121	374	2108	2942
				b	121	369	2130	3053
			i	m	121	374	2107	2939
				b	121	369	2129	3051
		Low	r	m	111	391	2100	2971
				b	111	383	2117	2961
			i	m	111	390	2100	2976
				b	111	383	2117	2962
	Adult	High	Regular	m	116	384	2095	2940
				b	116	367	2084	2971
			Irregular	m	116	384	2095	2940
				b	116	366	2085	2967
		Mid	r	m	109	390	2067	2960
				b	109	372	2059	2959
			i	m	109	390	2067	2957
				b	109	372	2059	2958
		Low	r	m	101	375	2062	3028
				b	101	378	2072	2992
			i	m	101	374	2064	3041
				b	101	378	2073	2994
	Senior	High	Regular	m	151	362	2057	2956
				b	151	373	2092	2949
			Irregular	m	151	362	2055	2957
				b	151	372	2092	2951
		Mid	r	m	129	372	2062	2935
				b	129	385	2058	2927
			i	m	129	372	2063	2946
				b	129	385	2058	2927
		Low	r	m	106	361	2020	2945
				b	106	401	2050	2951
			i	m	106	362	2012	2915
				b	106	401	2050	2946

Table E.11: F_1, F_2, F_3, and f_0 in **vowel [ɔ]** in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f_0 level	Irregularity	Breathiness	f_0 mean [Hz]	F_1 [Hz]	F_2 [Hz]	F_3 [Hz]
[ɔ] in [ɔ͡ɪ]	Young	High	Regular	Modal	129	546	923	2705
				Breathy	129	561	924	2695
			Irregular	m	128	546	924	2704
				b	128	564	927	2693
		Mid	r	m	119	546	911	2704
				b	119	553	925	2769
			i	m	118	546	911	2701
				b	118	552	923	2769
		Low	r	m	109	546	909	2708
				b	109	526	904	2734
			i	m	108	546	909	2705
				b	108	525	903	2731
	Adult	High	Regular	m	113	534	895	2639
				b	113	537	951	2625
			Irregular	m	113	537	897	2627
				b	113	532	947	2622
		Mid	r	m	106	507	869	2649
				b	105	509	936	2686
			i	m	105	509	870	2644
				b	105	509	936	2683
		Low	r	m	98	504	864	2618
				b	98	502	927	2605
			i	m	98	503	863	2623
				b	98	502	926	2608
	Senior	High	Regular	m	147	546	824	2605
				b	147	551	927	2582
			Irregular	m	146	541	837	2613
				b	146	547	923	2583
		Mid	r	m	124	558	963	2645
				b	124	469	885	2570
			i	m	124	547	952	2635
				b	124	468	884	2574
		Low	r	m	102	456	906	2590
				b	101	488	870	2596
			i	m	101	450	904	2594
				b	101	483	868	2586

Table E.12: F$_1$, F$_2$, F$_3$, and f$_0$ in **vowel [ʊ]** in the different voice profiles. Cf. Figure 7.1, Section 7.3.

Origin	Age class	f$_0$ level	Irregularity	Breathiness	f$_0$ mean [Hz]	F$_1$ [Hz]	F$_2$ [Hz]	F$_3$ [Hz]
[ʊ] in [aʊ̯]	YOUNG	High	Regular	Modal	131	366	801	2722
				Breathy	131	358	805	2790
			Irregular	m	131	366	801	2722
				b	131	357	805	2791
		Mid	r	m	121	363	804	2713
				b	121	360	810	2840
			i	m	121	363	804	2714
				b	121	361	811	2835
		Low	r	m	111	376	802	2740
				b	111	364	781	2798
			i	m	111	376	802	2741
				b	111	364	781	2799
	ADULT	High	Regular	m	116	351	782	2722
				b	116	330	803	2701
			Irregular	m	116	352	783	2711
				b	116	330	803	2698
		Mid	r	m	108	347	760	2643
				b	108	329	786	2628
			i	m	108	347	760	2644
				b	108	329	788	2628
		Low	r	m	101	342	769	2635
				b	101	327	779	2630
			i	m	101	342	769	2636
				b	101	327	779	2630
	SENIOR	High	Regular	m	151	314	749	2373
				b	151	341	756	2412
			Irregular	m	151	314	747	2378
				b	151	341	755	2406
		Mid	r	m	129	312	777	2355
				b	128	356	743	2456
			i	m	129	311	768	2337
				b	129	357	744	2460
		Low	r	m	106	303	797	2539
				b	106	363	744	2491
			i	m	106	304	799	2538
				b	106	362	743	2490

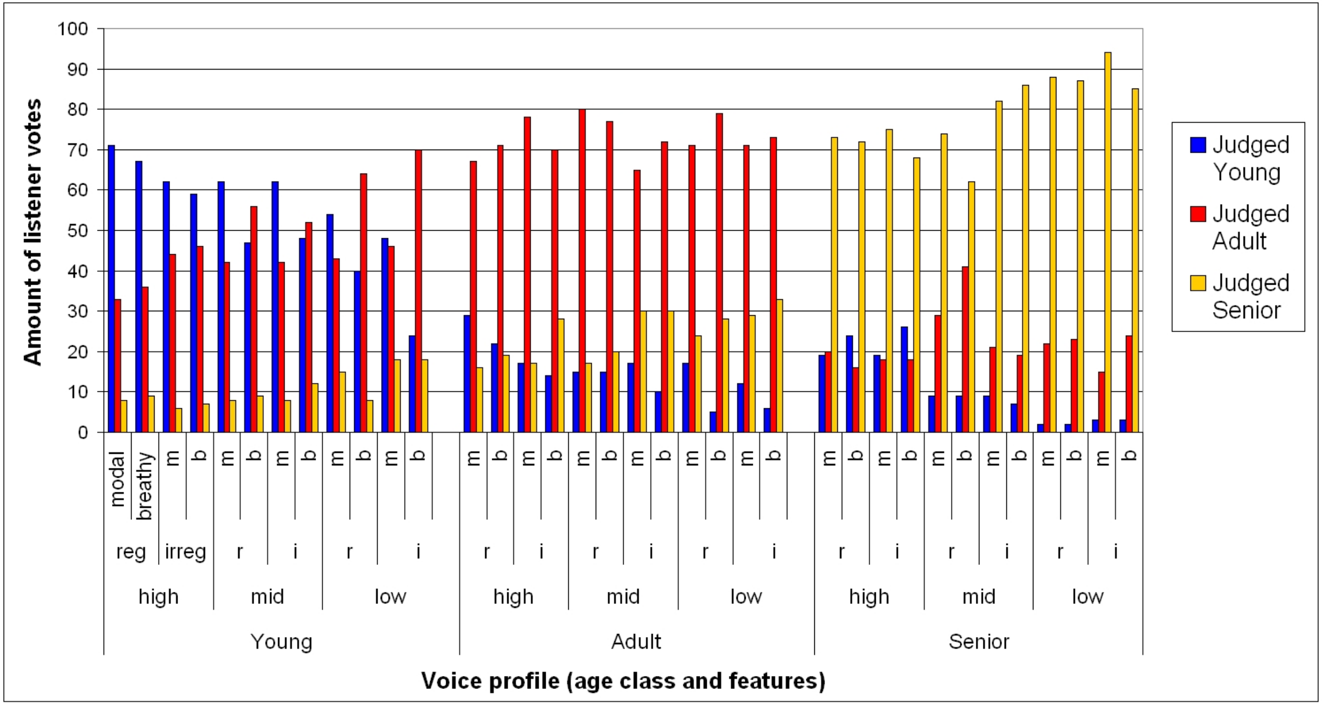
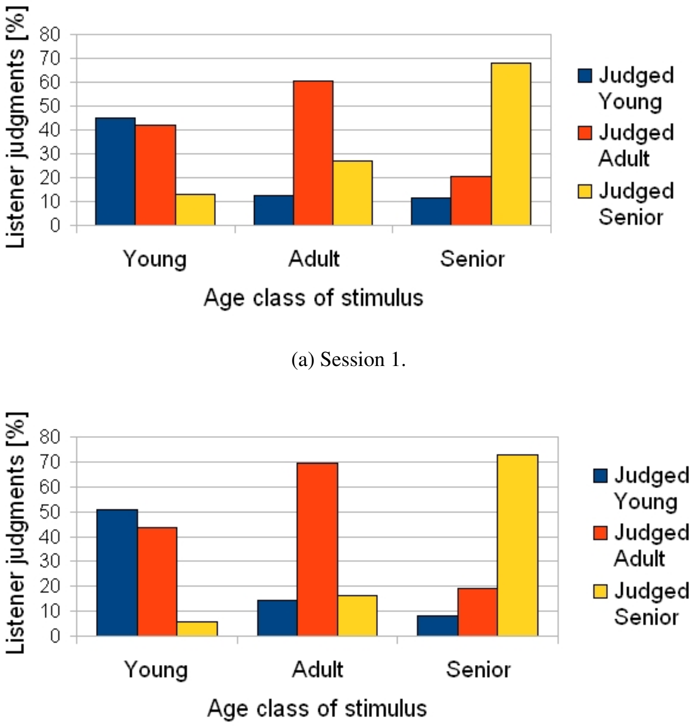


Figure E.1: Number of listener judgments for each voice profile. Based on 4036 judgments from 28 participants. Further details are provided in Table E.13.

Table E.13: Detailed listing of the number of listener judgments for each voice profile. Based on 4036 judgments from 28 participants and visualized in Figure E.1. Each voice profile is represented by 2 stimuli with different diphthong pairings ([a͜ɪ a͜ʊ], [a͜ʊ ɔ͜ɪ]). Each stimulus was rated 56 times, thus 112 ratings for each corresponding voice profile. See also Section 7.3.2.

Age class	**f**$_0$	**Irregularity**	**Breathiness**	**Judged** YOUNG	**Judged** ADULT	**Judged** SENIOR
YOUNG	High	Regular (r)	Modal (m)	71	33	8
			Breathy (b)	67	36	9
		Irregular (i)	m	62	44	6
			b	59	46	7
	Mid	r	m	62	42	8
			b	47	56	9
		i	m	62	42	8
			b	48	52	12
	Low	r	m	54	43	15
			b	40	64	8
		i	m	48	46	18
			b	24	70	18
ADULT	High	Regular (r)	Modal (m)	29	67	16
			Breathy (b)	22	71	19
		Irregular (i)	m	17	78	17
			b	14	70	28
	Mid	r	m	15	80	17
			b	15	77	20
		i	m	17	65	30
			b	10	72	30
	Low	r	m	17	71	24
			b	5	79	28
		i	m	12	71	29
			b	6	73	33
SENIOR	High	Regular (r)	Modal (m)	19	20	73
			Breathy (b)	24	16	72
		Irregular (i)	m	19	18	75
			b	26	18	68
	Mid	r	m	9	29	74
			b	9	41	62
		i	m	9	21	82
			b	7	19	86
	Low	r	m	2	22	88
			b	2	23	87
		i	m	3	15	94
			b	3	24	85



(a) Session 1.

(b) Session 2.

Figure E.2: Judgments of 28 listeners, session 1 vs. session 2, based on 2016 judgments for each session. Cf. also Table E.14.

Table E.14: Confusion matrices showing the age judgments of the 28 participants of the forced-choice classification test in **session 1** vs. **session 2**. The numbers are based on 2016 votes for each session. The chance level is at 33.33 % since the subjects were allowed to choose between three age classes. Overall, **58.04 %** of the samples were classified correctly in session 1, **64.19 %** in session 2. Cf. also Figure E.2.

Age class of	Listener judgments session 1				Listener judgments session 2			
stimulus	YOUNG	ADULT	SENIOR	Total	YOUNG	ADULT	SENIOR	Total
YOUNG	**45.24 %**	41.96 %	12.80 %	100 %	**50.60 %**	43.45 %	5.95 %	100 %
ADULT	12.50 %	**60.71 %**	26.79 %	100 %	14.14 %	**69.35 %**	16.52 %	100 %
SENIOR	11.31 %	20.54 %	**68.15 %**	100 %	8.33 %	19.05 %	**72.62 %**	100 %

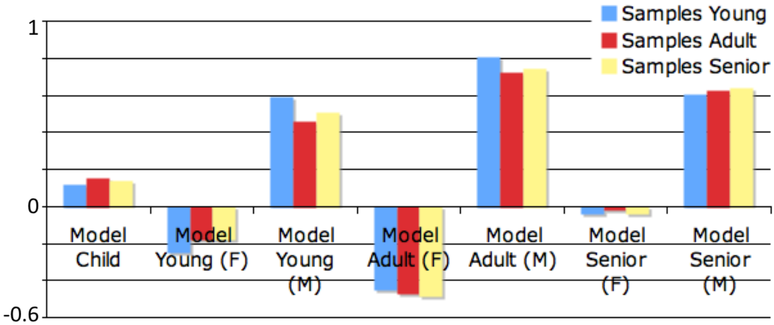


Figure E.3: Rating scores of all 7 age models of the automatic age classification system. While the age classes are often misrecognized, the classifier at least reliably picks the correct gender. (F) female, (M) male. Graph provided by C. Müller. See also the resulting age class decisions shown in the confusion matrix in Figure E.15.

Table E.15: Confusion matrix illustrating the classification accuracy of an automatic age classification system (cf. Feld, 2011, Müller, 2005). The chance level is at 14.28 % since it represents a 7-class problem for the system, having been trained on 7 age class models (compare Figure E.3, see also Section 7.4.3). However, only *male* YOUNG, ADULT, and SENIOR numbers are of interest here. The overall accuracy is at a level of **29.17 %** of correctly classified samples. For comparison, in a different experiment with this system, an accuracy of roughly 60 % was obtained on an independent evaluation of natural speech samples (cf. also Müller, 2006, for further comparisons cf. also numbers in Schötz, 2006: 53ff).

Age class	Model YOUNG	Model ADULT	Model SENIOR	All classes
Sample YOUNG	**16.67** %	79.16 %	4.17 %	100 %
Sample ADULT	0.00 %	**54.17 %**	45.83 %	100 %
Sample SENIOR	0.00 %	83.33 %	**16.67** %	100 %

Appendix F

Supplementary material from the Saxon accent experiment

The material provided in this part of the Appendix is referenced to at different points in the regional accent experiment, presented in Chapters 10 and 11. Details are given on recorded words and vocal tract configurations of the newly created vowels (Section F.1), settings for word synthesis (Section F.2) and details of the statistical analyses of the accent perception test (Section F.3).

F.1 Lists of recorded words and vocal tract configurations

Tables F.1 and F.2 list the corpus that was recorded for the regional accent experiment. Only parts of it were used in this thesis, as described in Section 10.2.1. An impression of the occurrence frequency of the used words is provided in Table F.3. Table F.4 provides an overview of the supraglottal configurations of the Saxon and Standard High German vowels.

F.2 Word synthesis: Dominance settings and gestural scores

Table F.5 lists the settings of different versions of /f/ and /l/, explored and used for word synthesis. The following gestural scores illustrate how the words for the perception test are generated: The regionally accented versions are built by exchanging the first vowel in each word from Standard High German (SHG) to Saxon-accented (not depicted). The words illustrate different usages of the same vocal tract configurations, e.g. producing [s] and [z] from one configuration, depending on the degree of glottal opening. Furthermore, score (c) illustrates the usage of two slightly different /yː/ configurations to optimize the sound of the initial consonants, although the vowel is acoustically hidden most of that time. Each score covers a duration of 0.8 s.

Table F.1: List of recorded words, sorted alphabetically, part 1. 18 two-syllable verb infinitives, marked in bold face, are used for the perception test presented in Chapter 11.

aber
Adler
Apfel
Arzt
Bagger
Bananen
Beere
beten
Biene
bieten
Birne
blöde
blöken
Blöße
blühen
Blume
Blumen
Bluse
Blüte
bluten
Blüten
Boden
Bogen
Böhmen
bohren
Boote
böse
Brötchen
Brügger
Brühe
brühen
brüten
Bube
Buche
buchen
Bühne
Chemikerin
die Toten
Diebe
Döschen
Dose
dösen
drehen
drohen
dröhnen
Drüse
Düne
Dusche
Düse
Eber
Ehe
Einbahnstraßenschild

Eis
Eishockeyspieler
Elektriker
Ente
Erdbeere
Erdbeeren
Fernglas
fiepen
flehen
Fliege
fliegen
fliehen
Fliese
fließen
Flöhe
Flöte
fluchen
fluten
Frieden
frieren
Fuchs
Fuge
fügen
fühlen
Fuhre
führen
Fußballspieler
geben
gehen
Getreide
gießen
Glas
Glibber
Glühbirne
glühen
Größe
Grube
Grüße
grüßen
Gürtel
Güte
Hamburger
Hammer
Handy
Headset
heben
Hefe
hegen
Helicopter
Hirsch
Höfe
Höhe

Höhle
holen
hören
Höschen
Hose
Hund
Hüne
Hupe
hupen
Hure
husten
Hüte
hüten
ihre
Jäger
Joghurtbecher
Jute
Kaffeemaschine
Karotten
Karton
Käse
Katze
Kehle
Kehre
kehren
Kirsche
klonen
Klösschen
Kohle
Kooge
kriechen
Kriege
kriegen
Krokodil
Kröte
Kuchen
Kufe
Kufen
Kühe
Kuhle
Kühle
kühlen
Küken
küren
leben
Leber
Leere
legen
Lehne
lehnen
lehren
lesen

Liebe
lieben
lieber
Liege
liegen
loben
Loge
löhnen
Lose
losen
lösen
löten
Lotse
lotsen
Löwe
Lüge
lügen
Luke
lumen
Messer
Metzger
Miene
Miete
mieten
Mikroskop
Mobber
Mode
mögen
Möhre
Monologe
Möwe
Mühe
mühen
Mühle
neben
nehmen
Niere
Niete
nieten
Nische
Nöte
Noten
ober
öde
Ofen
Ohren
Öle
ölen
Orangen
Öse
Pfanne
Pflaster

Pflege
Pieke
piepen
Pommes Frittes
Pose
Probe
proben
prüde
prüfen
pusten
Rad
Rasenmäher
Rebe
Rede
reden
rege
Regenwurm
Riese
Röhre
Röschen
Rose
rote
Röte
Rübe
Rüde
rufen
Rüge
rügen
Ruhe
ruhen
rühmen
rühren
Salat
Sanduhr
Schere
schieben
Schiene
schienen
schießen
Schlehe
schließen
Schmetterling
Schmiede
schmieden
schmieren
Schnecke
Schneegestöber
schnüren
schönen
Schornsteinfeger
Schuber
schwören

Table F.2: List of recorded words, sorted alphabetically, part 2. 18 two-syllable verb infinitives, marked in bold face, are used for the perception test presented in Chapter 11.

Schwüle	Spiegelei	stöhnen	**sühnen**	töten	Vene	**wohnen**	ziemen
Seele	spielen	Stopschild	Telefon	Tresen	Ventilator	wühlen	Zitrone
sehen	spröde	stören	Theke	Triebe	verpönen	Wüste	**zoomen**
Sekretärin	sprühen	Stöße	These	trübe	Wehe	wüten	Zug
sieben	Spule	stoßen	thronen	Truhe	wehen	Zahnarzt	Züge
siechen	Spüle	streben	Tierarzt	Truthahn	Wesen	Zahnbürste	
sieden	spülen	strömen	Toaster	Tüte	Wiege	Zange	
siegen	spüren	Stufe	toben	**tuten**	wiegen	Zehe	
siezen	stehen	Suche	Töne	üben	Wiese	Zehen	
Sohle	stieben	suchen	**tosen**	über	Windrad	Ziege	
Soße	stieren	Sühne	töte	unter	Woge	ziehen	

Table F.3: Word occurrence frequencies (words per million) of the 18 two-syllable words used in the perception test presented in Chapter 11. The list is sorted by target vowels (first vowel in each word), including case sensitive variations of the words if applicable. The values are extracted from Celex (Baayen et al., 1995). Since Celex is reported to be "relatively ill-balanced" (cf. http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html) and some words show zero occurrences per million in Celex, numbers from Google books Ngram Viewer (http://books.google.com/ngrams) are supplied as a second source of information, based on "a lot of" scanned books "predominantly in the German language" (cf. http://books.google.com/ngrams/info) from the years 1998 to 2008, queried on May 6, 2013. Word frequencies of the most frequent words in German, according to http://wortschatz.uni-leipzig.de, are shown for comparison.

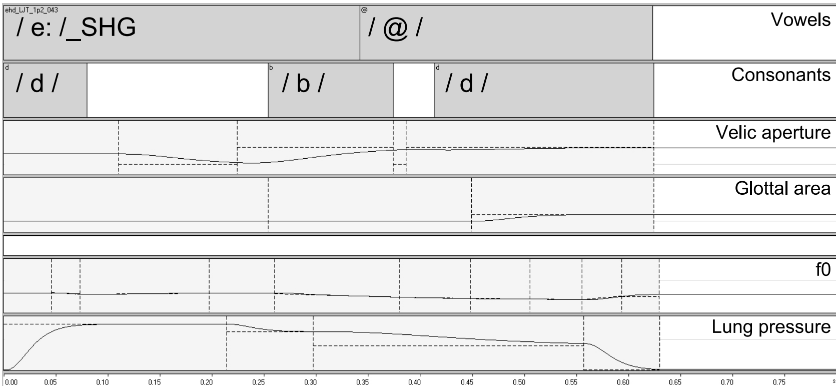
Celex		**Google books Ngram Viewer**	
bieten	194	bieten + Bieten	60.44
schienen	6	schienen + Schienen	12.00
sieden	0	sieden + Sieden	0.38
leben + Leben	873	leben + Leben	387.82
lehnen	67	lehnen + Lehnen	4.20
nehmen	818	nehmen + Nehmen	143.30
loben	21	loben + Loben	2.66
tosen	1	tosen + Tosen	0.16
wohnen	79	wohnen + Wohnen	17.87
hupen	1	hupen + Hupen	0.23
tuten	2	tuten + Tuten	0.12
zoomen	0	zoomen + Zoomen	0.20
bloeken	1	blöken + Blöken	0.08
loesen	147	lösen + Lösen + loesen + Loesen	41.70
schoenen	53	schönen + Schönen + schoenen + Schoenen	25.22
fuegen	39	fügen + Fügen + fuegen + Fuegen	7.21
gruessen	17	grüßen + Grüßen + gruessen + Gruessen + grüssen + Grüssen	4.50
suehnen	2	sühnen + Sühnen + suehnen + Suehnen	0.28
Most frequent words in German for comparison			
der	36577	der + Der	29470
die	n.a.	die + Die	27808
und	25287	und + Und	21851
in	18536	in + In	15635
den	n.a.	den + Den	8427
von	4693	von + Von	8694

Table F.4: Vocal tract configurations of the accented vowels, generated by the formant optimization algorithm (see Section 4.2.4). The meaning of the parameters is explained in Section 3.1.1, Table 3.1. All parameters have been manipulated to generate the Saxon (*Sax*) and Standard High German (*SHG*) versions of the vowels, except for tongue center radius $TCR_{X,Y}$, tongue side elevation TS_{1-4} and minimal cross-sectional area MA_{1-3}. All parameters of /øː/$_{SHG}$ and /øː/$_{Sax}$ have a dominance value of 100, in the other vowels it is set to 0. However, use on the gestural score asserts a dominance value of 100 to gestures placed on the vowel tier.

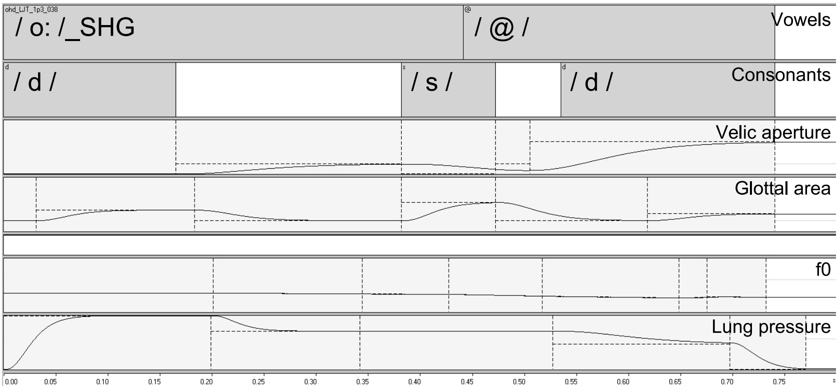
Parameter name	/øː/$_{SHG}$	/øː/$_{Sax}$	/eː/$_{SHG}$	/eː/$_{Sax}$	/iː/$_{SHG}$	/iː/$_{Sax}$	/oː/$_{SHG}$	/oː/$_{Sax}$	/uː/$_{SHG}$	/uː/$_{Sax}$	/yː/$_{SHG}$	/yː/$_{Sax}$
HX	0.0239	0.0224	0.2024	0.2044	0.9946	0.9947	0.6474	0.6453	0.9906	0.9841	0.5299	0.5204
HY	-5.3694	-5.367	-4.3163	-4.3185	-5.1075	-5.1209	-5.3236	-5.3192	-5.3429	-5.3498	-4.4378	-4.4584
JX	-0.3925	-0.3168	-0.3016	-0.226	-0.3649	-0.4198	-0.1568	-0.2569	-0.125	-0.1416	-0.2527	-0.1794
JY	-1.3032	-1.3303	-1.2311	-1.2326	-1.2061	-1.2086	-1.3273	-1.2699	-1.2687	-1.239	-1.3573	-1.3413
JA	-0.1343	-0.1246	-0.1099	-0.1204	-0.0641	-0.0697	-0.1092	-0.0952	-0.088	-0.1007	-0.1356	-0.122
LP	0.7116	0.7241	0.0537	-0.0527	0.0662	0.0927	0.8558	0.9646	0.926	0.8642	0.8602	0.7615
LH	0.2304	0.6018	0.5799	0.6545	0.3996	0.4837	0.1987	0.1386	0.0022	0.0548	0.2252	0.3895
VEL	0.0032	0.0001	0.042	0.0438	0.0015	0.0009	0.0359	0.0253	0.0079	0	0.0195	0.025
TCX	1.7509	1.4099	2.4913	2.0486	2.1094	1.9689	-0.1348	0.4728	0.3875	0.462	2.3308	2.1336
TCY	-0.973	-1.2362	-0.8594	-1.0907	-0.9221	-1.1443	-0.8155	-1.3722	-0.712	-0.9581	-0.5799	-0.8128
TCRX	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
TCRY	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8
TTX	4.287	4.3376	4.83	4.5933	4.4272	4.3429	2.2358	3.4878	2.8185	3.349	4.7184	4.5727
TTY	-1.1716	-1.1521	-0.8172	-0.8652	-0.6247	-0.6885	-1.4621	-0.885	-0.5732	-0.3438	-1.0657	-1.1503
TBX	3.6823	3.5435	3.9202	3.8201	3.1744	2.9948	2.1388	1.8086	2.5205	2.819	2.5594	3.6796
TBY	-0.1987	0.9427	0.2587	0.4977	1.3495	1.3872	0.4109	0.2952	1.0812	0.8913	1.7385	0.7184
TRX	-0.6589	-0.8203	-0.0479	-0.4984	-0.1616	-0.1594	-2.019	-2.1428	-1.5625	-1.2744	-1.5316	-0.5599
TRY	-2.1111	-1.8314	-0.9971	-1.0907	-2.6015	-2.9133	-2.8535	-2.8837	-2.8601	-2.9839	-1.8952	-1.2877
TS1	1	1	0.9025	0.9025	1	1	0.8025	0.8025	0.99	0.99	1	1
TS2	0.1875	0.1875	0.1075	0.1075	0.26	0.26	-0.4775	-0.4775	-0.0025	-0.0025	0.3875	0.3875
TS3	0.35	0.35	-0.015	-0.015	0.15	0.15	0.1025	0.1025	0.0625	0.0625	0.015	0.015
TS4	0.3125	0.3125	0.0025	0.0025	-0.01	-0.01	-0.0525	-0.0525	-0.285	-0.285	-0.0475	-0.0475
MA1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
MA2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
MA3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2

Table F.5: Articulatory parameters (*VT param.*) and dominance values of different versions of the labiodental fricative /f/ and the alveolar lateral-approximant /l/, introduced in Section 11.1.2, Figures 11.2 and 11.3. Only those dominance values were changed that originally were not set to 100 (printed in bold face) and, in the case of /l/, those that seemed relevant to the production or avoiding of 'dark', velarized /l/. The dominance value of HX was not manipulated in /f/ because it seemed non-critical for the perceived distortions. Descriptions of the vocal tract parameters are given in Section 3.1.1, Table 3.1. Please note that the geometric target definition (*Value* column) is identical for all versions of /f/ and all versions of /l/, respectively, i.e. when produced in isolation they all sound the same. Only the dominance of each sound against *other* sounds is manipulated.

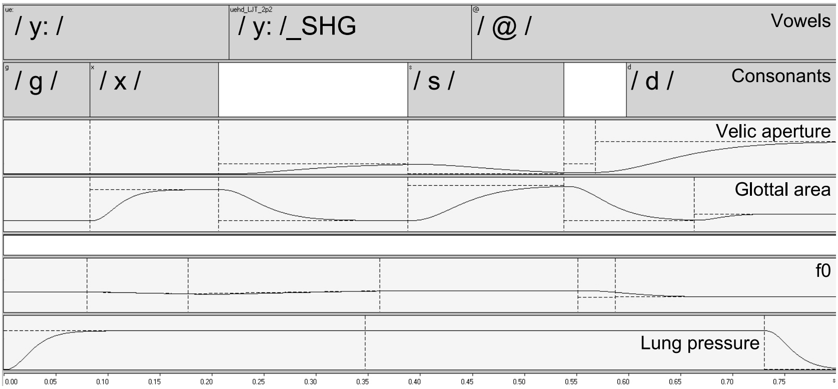
VT		Dominance values of ...				VT		Dominance values of ...			
param.	Value	f	f40	f60	f100-TC40	param.	Value	l	l100	l80	l70
HX	0.0000	15	15	15	15	**HX**	0.3763	**15**	**100**	**15**	**70**
HY	-5.2161	100	100	100	100	HY	-4.9008	100	100	100	100
JX	-0.4573	100	100	100	100	JX	-0.1419	100	100	100	100
JY	-1.2000	100	100	100	100	**JY**	-1.2000	**78**	**100**	**80**	**78**
JA	-0.0355	100	100	100	100	**JA**	-0.0876	**80**	**100**	**80**	**80**
LP	-0.0534	100	100	100	100	**LP**	0.0450	**39**	**100**	**80**	**39**
LH	0.2628	100	100	100	100	**LH**	0.4943	**46**	**100**	**80**	**46**
VEL	0.1323	100	100	100	100	VEL	0.0000	100	100	100	100
TCX	0.2847	**0**	**40**	**60**	**40**	**TCX**	0.6927	**23**	**100**	**80**	**70**
TCY	-2.2113	**10**	**40**	**60**	**40**	**TCY**	-2.0629	**3**	**100**	**80**	**70**
TCRX	1.8000	100	100	100	100	TCRX	1.8000	100	100	100	100
TCRY	1.8000	100	100	100	100	TCRY	1.8000	100	100	100	100
TTX	3.5677	**0**	**40**	**60**	**100**	TTX	4.5000	100	100	100	100
TTY	-1.3766	**30**	**40**	**60**	**100**	TTY	-0.2452	100	100	100	100
TBX	1.7871	**10**	**40**	**60**	**100**	**TBX**	2.1210	**51**	**100**	**80**	**70**
TBY	-0.3565	**10**	**40**	**60**	**100**	**TBY**	0.1258	**18**	**100**	**80**	**70**
TRX	-2.1637	**10**	**40**	**60**	**100**	TRX	-1.6073	0	0	0	0
TRY	-2.5081	100	100	100	100	TRY	-2.7492	100	100	100	100
TS1	0.2300	**30**	**40**	**60**	**100**	**TS1**	0.5400	**75**	**100**	**75**	**75**
TS2	0.1100	**30**	**40**	**60**	**100**	**TS2**	-0.0800	**42**	**100**	**42**	**70**
TS3	-0.2000	**10**	**40**	**60**	**100**	TS3	-0.9700	100	100	100	100
TS4	-0.3900	**10**	**40**	**60**	**100**	TS4	-0.2400	100	100	100	100
MA1	0.2000	100	100	100	100	MA1	0.1500	100	100	100	100
MA2	0.1500	100	100	100	100	MA2	0.1500	100	100	100	100
MA3	0.1500	100	100	100	100	MA3	0.0000	100	100	100	100



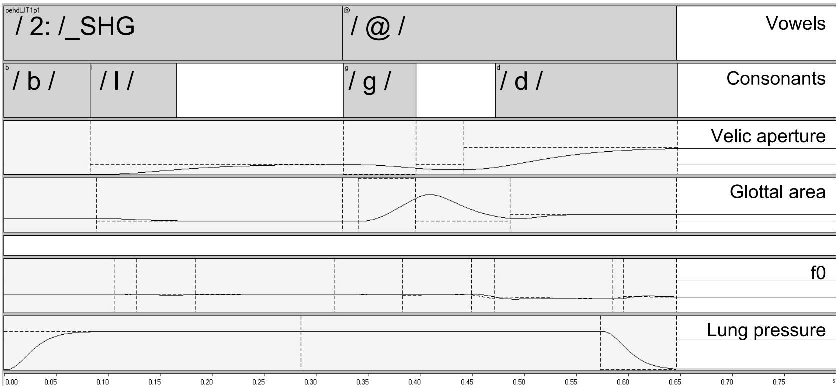
(a) nehmen$_{SHG}$



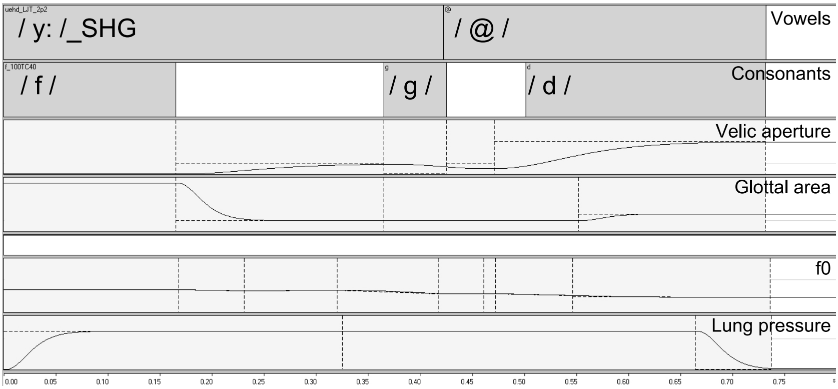
(b) tosen$_{SHG}$



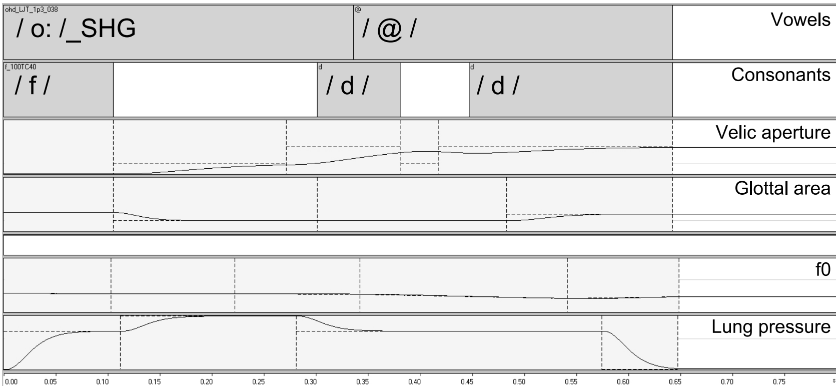
(c) grüßen$_{SHG}$



(d) blöken$_{SHG}$



(e) fügen$_{SHG}$



(f) wohnen$_{SHG}$

F.3 Statistical details of the perception experiment

Tables F.6 to F.11 and Figures F.1 to F.2 provide details of the statistical analysis of the accent perception experiment presented in Chapter 11.

Table F.6: Mean accent ratings (least square means and standard deviations) in each test condition (voice), across all stimuli, by intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German). See also Figure 11.8.

Voice	**Intended accent**	
	Standard High German	Saxon
Synthetic	2,66 (0,19)	4,97 (0,19)
Human	2,25 (0,17)	5,40 (0,17)

Table F.7: Mean accent ratings (least square means and standard deviations) of the synthetic voice stimuli, by listeners' language background and intended accent. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German). See also Figure 11.8c.

Listeners' language background	**Intended accent**	
	Standard High German	Saxon
East Central German	2,63 (0,23)	5,10 (0,23)
Not East Central German	2,69 (0,19)	4,84 (0,19)

Table F.8: Mean accent ratings (least square means) for each carrier vowel of both voices, by intended accent (SHG = Standard High German). Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German). See also Figure 11.9.

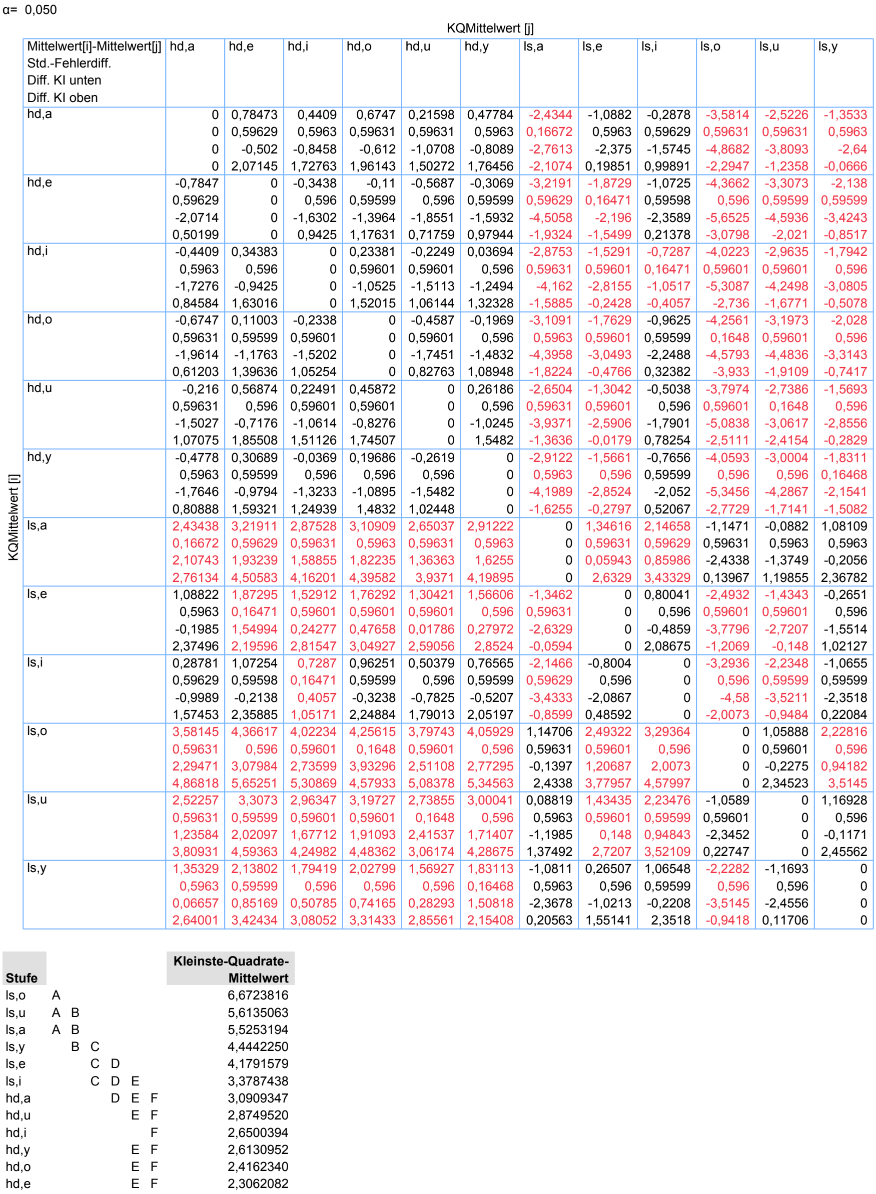
Vowel	**Intended accent**		**Vowel**	**Intended accent**	
	SHG	*Saxon*		*SHG*	*Saxon*
Synthetic voice			Human voice		
/øː/	3.09 (0.43)	5.53 (0.43)	/øː/	2.01 (0.39)	5.37 (0.39)
/eː/	2.31 (0.43)	4.18 (0.43)	/eː/	2.34 (0.39)	3.75 (0.39)
/iː/	2.65 (0.43)	3.38 (0.43)	/iː/	2.76 (0.39)	4.41 (0.39)
/oː/	2.42 (0.43)	6.67 (0.43)	/oː/	2.09 (0.39)	6.62 (0.39)
/uː/	2.87 (0.43)	5.61 (0.43)	/uː/	2.02 (0.39)	6.27 (0.39)
/yː/	2.61 (0.43)	4.44 (0.43)	/yː/	2.28 (0.39)	6.00 (0.39)

Table F.9: Vowel contrasts between the vowel pairs /vowel/$_{SHG}$ and /vowel/$_{Sax}$. Consolidated rating scale: 1 = Not Saxon (Very High German), 7 = Very Saxon (Not High German). See also Figure 11.9.

Vowel (pair)	/øː/	/eː/	/iː/	/oː/	/uː/	/yː/
Synthetic voice						
t(1)	-14.6	-11.37	-4.424	-25.83	-16.62	-11.12
p> $\|t\|$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Human voice						
t(1)	-13.12	-5.526	-6.446	-17.66	-16.55	-14.53
p> $\|t\|$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table F.10: Behavior of the carrier words regarding perceived accent. Effects within each group of stimuli which carry the same target vowel, tested for intended accent (intacc) and lexeme. Significant effects (p<0.05) are marked in bold face. See also Figures 11.10 and 11.11.

Vowel	Effect	Synthetic voice	Human voice
/øː/	intacc	**F(1, 217.3)=233.781, p<0.0001**	**F(1, 67.84)=154.5142, p<0.0001**
	lexem	**F(2, 232.9)=12.4252, p<0.0001**	**F(2, 75.92)=8.2094, p=0.0006**
	intacc*lexem	**F(2, 258.7)=5.981, p=0.0029**	**F(2, 87.84)=3.6325, p=0.0305**
/eː/	intacc	**F(1, 217.2)=112.9241, p<0.0001**	**F(1, 66.27)=16.4595, p=0.0001**
	lexem	**F(2, 227.9)=17.7801, p<0.0001**	F(2, 74.6)=2.5193, p=0.0873
	intacc*lexem	F(2, 249.4)=1.7249, p=0.1803	F(2, 87.34)=1.4741, p=0.2346
/iː/	intacc	**F(1, 215.9)=8.7781, p=0.0034**	**F(1, 67.48)=20.6823, p<0.0001**
	lexem	**F(2, 226)=24.2897, p<0.0001**	**F(2, 76.84)=18.6491, p<0.0001**
	intacc*lexem	F(2, 246.6)=0.2868, p=0.7509	F(2, 88.98)=1.4056, p=0.2506
/oː/	intacc	**F(1, 216.2)=972.4616, p<0.0001**	**F(1, 67.32)=553.9144, p<0.0001**
	lexem	**F(2, 236.5)=13.8089, p<0.0001**	F(2, 74.59)=2.5078, p=0.0883
	intacc*lexem	**F(2, 272.7)=18.6528, p<0.0001**	F(2, 85.98)=2.6559, p=0.076
/uː/	intacc	**F(1, 217.1)=216.8455, p<0.0001**	**F(1, 67.3)=278.9698, p<0.0001**
	lexem	**F(2, 233.6)=41.835, p<0.0001**	F(2, 74.07)=0.804, p=0.4514
	intacc*lexem	F(2, 265.5)=0.4422, p=0.6431	**F(2, 85.04)=8.6323, p=0.0004**
/yː/	intacc	**F(1, 217.2)=93.8869, p<0.0001**	**F(1, 67.56)=258.6106, p<0.0001**
	lexem	**F(2, 235.8)=24.1724, p<0.0001**	**F(2, 74.12)=3.4276, p=0.0377**
	intacc*lexem	F(2, 266)=0.9692, p=0.3807	F(2, 84.57)=1.6455, p=0.199



α= 0,050

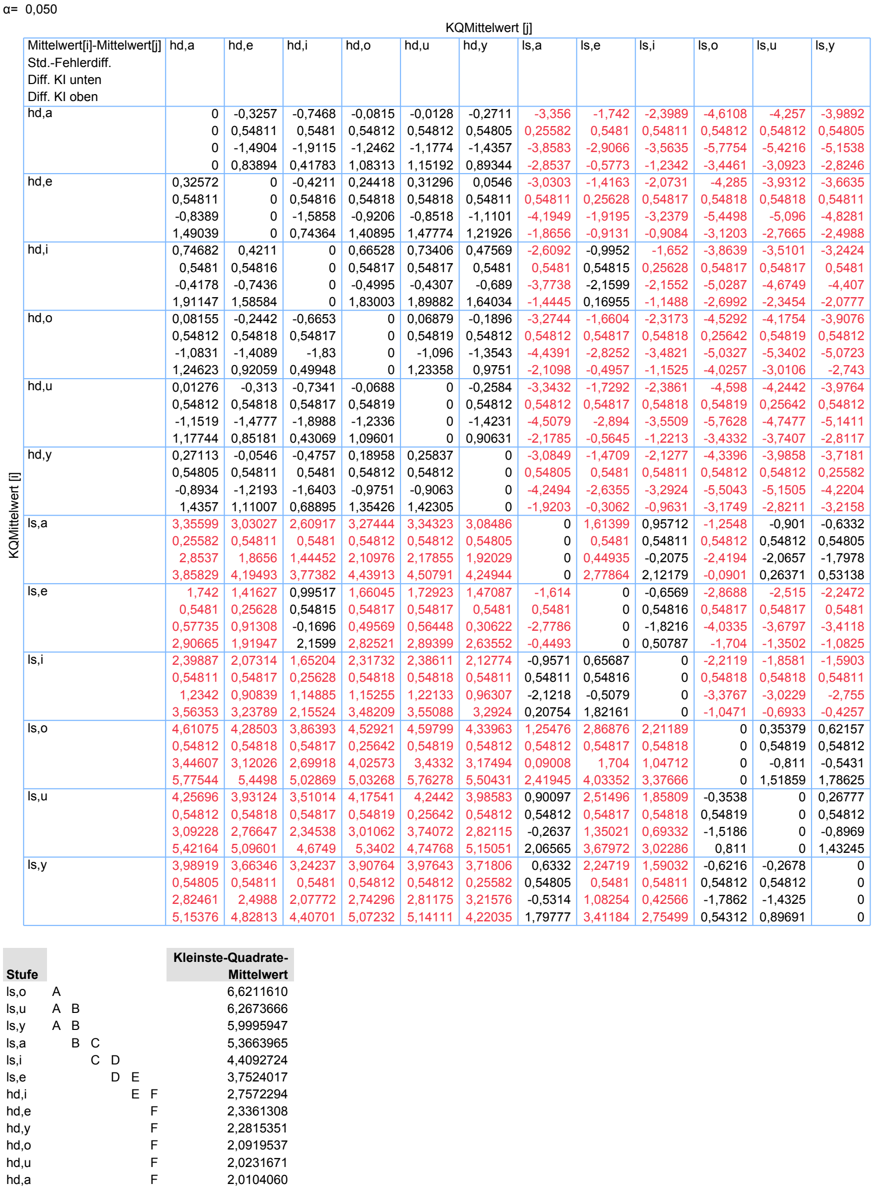
KQMittelwert [j]

KQMittelwert [i]

Mittelwert[i]-Mittelwert[j] Std.-Fehlerdiff. Diff. KI unten Diff. KI oben	hd,a	hd,e	hd,i	hd,o	hd,u	hd,y	ls,a	ls,e	ls,i	ls,o	ls,u	ls,y
hd,a	0	0,78473	0,4409	0,6747	0,21598	0,47784	-2,4344	-1,0882	-0,2878	-3,5814	-2,5226	-1,3533
	0	0,59629	0,5963	0,59631	0,59631	0,5963	0,16672	0,5963	0,59629	0,59631	0,59631	0,5963
	0	-0,502	-0,8458	-0,612	-1,0708	-0,8089	-2,7613	-2,375	-1,5745	-4,8682	-3,8093	-2,64
	0	2,07145	1,72763	1,96143	1,50272	1,76456	-2,1074	0,19851	0,99891	-2,2947	-1,2358	-0,0666
hd,e	-0,7847	0	-0,3438	-0,11	-0,5687	-0,3069	-3,2191	-1,8729	-1,0725	-4,3662	-3,3073	-2,138
	0,59629	0	0,596	0,59599	0,596	0,59599	0,59629	0,16471	0,59598	0,596	0,59599	0,59599
	-2,0714	0	-1,6302	-1,3964	-1,8551	-1,5932	-4,5058	-2,196	-2,3589	-5,6525	-4,5936	-3,4243
	0,50199	0	0,9425	1,17631	0,71759	0,97944	-1,9324	-1,5499	0,21378	-3,0798	-2,021	-0,8517
hd,i	-0,4409	0,34383	0	0,23381	-0,2249	0,03694	-2,8753	-1,5291	-0,7287	-4,0223	-2,9635	-1,7942
	0,5963	0,596	0	0,59601	0,59601	0,596	0,59631	0,59601	0,16471	0,59601	0,59601	0,596
	-1,7276	-0,9425	0	-1,0525	-1,5113	-1,2494	-4,162	-2,8155	-1,0517	-5,3087	-4,2498	-3,0805
	0,84584	1,63016	0	1,52015	1,06144	1,32328	-1,5885	-0,2428	-0,4057	-2,736	-1,6771	-0,5078
hd,o	-0,6747	0,11003	-0,2338	0	-0,4587	-0,1969	-3,1091	-1,7629	-0,9625	-4,2561	-3,1973	-2,028
	0,59631	0,59599	0,59601	0	0,59601	0,596	0,5963	0,59601	0,59599	0,1648	0,59601	0,596
	-1,9614	-1,1763	-1,5202	0	-1,7451	-1,4832	-4,3958	-3,0493	-2,2488	-4,5793	-4,4836	-3,3143
	0,61203	1,39636	1,05254	0	0,82763	1,08948	-1,8224	-0,4766	0,32382	-3,933	-1,9109	-0,7417
hd,u	-0,216	0,56874	0,22491	0,45872	0	0,26186	-2,6504	-1,3042	-0,5038	-3,7974	-2,7386	-1,5693
	0,59631	0,596	0,59601	0,59601	0	0,596	0,59631	0,59601	0,596	0,59601	0,1648	0,596
	-1,5027	-0,7176	-1,0614	-0,8276	0	-1,0245	-3,9371	-2,5906	-1,7901	-5,0838	-3,0617	-2,8556
	1,07075	1,85508	1,51126	1,74507	0	1,5482	-1,3636	-0,0179	0,78254	-2,5111	-2,4154	-0,2829
hd,y	-0,4778	0,30689	-0,0369	0,19686	-0,2619	0	-2,9122	-1,5661	-0,7656	-4,0593	-3,0004	-1,8311
	0,5963	0,59599	0,596	0,596	0,596	0	0,5963	0,596	0,59599	0,596	0,596	0,16468
	-1,7646	-0,9794	-1,3233	-1,0895	-1,5482	0	-4,1989	-2,8524	-2,052	-5,3456	-4,2867	-2,1541
	0,80888	1,59321	1,24939	1,4832	1,02448	0	-1,6255	-0,2797	0,52067	-2,7729	-1,7141	-1,5082
ls,a	2,43438	3,21911	2,87528	3,10909	2,65037	2,91222	0	1,34616	2,14658	-1,1471	-0,0882	1,08109
	0,16672	0,59629	0,59631	0,5963	0,59631	0,5963	0	0,59631	0,59629	0,59631	0,5963	0,5963
	2,10743	1,93239	1,58855	1,82235	1,36363	1,6255	0	0,05943	0,85986	-2,4338	-1,3749	-0,2056
	2,76134	4,50583	4,16201	4,39582	3,9371	4,19895	0	2,6329	3,43329	0,13967	1,19855	2,36782
ls,e	1,08822	1,87295	1,52912	1,76292	1,30421	1,56606	-1,3462	0	0,80041	-2,4932	-1,4343	-0,2651
	0,5963	0,16471	0,59601	0,59601	0,59601	0,596	0,59631	0	0,596	0,59601	0,59601	0,596
	-0,1985	1,54994	0,24277	0,47658	0,01786	0,27972	-2,6329	0	-0,4859	-3,7796	-2,7207	-1,5514
	2,37496	2,19596	2,81547	3,04927	2,59056	2,8524	-0,0594	0	2,08675	-1,2069	-0,148	1,02127
ls,i	0,28781	1,07254	0,7287	0,96251	0,50379	0,76565	-2,1466	-0,8004	0	-3,2936	-2,2348	-1,0655
	0,59629	0,59598	0,16471	0,59599	0,596	0,59599	0,59629	0,596	0	0,596	0,59599	0,59599
	-0,9989	-0,2138	0,4057	-0,3238	-0,7825	-0,5207	-3,4333	-2,0867	0	-4,58	-3,5211	-2,3518
	1,57453	2,35885	1,05171	2,24884	1,79013	2,05197	-0,8599	0,48592	0	-2,0073	-0,9484	0,22084
ls,o	3,58145	4,36617	4,02234	4,25615	3,79743	4,05929	1,14706	2,49322	3,29364	0	1,05888	2,22816
	0,59631	0,596	0,59601	0,1648	0,59601	0,596	0,59631	0,59601	0,596	0	0,59601	0,596
	2,29471	3,07984	2,73599	3,93296	2,51108	2,77295	-0,1397	1,20687	2,0073	0	-0,2275	0,94182
	4,86818	5,65251	5,30869	4,57933	5,08378	5,34563	2,4338	3,77957	4,57997	0	2,34523	3,5145
ls,u	2,52257	3,3073	2,96347	3,19727	2,73855	3,00041	0,08819	1,43435	2,23476	-1,0589	0	1,16928
	0,59631	0,59599	0,59601	0,59601	0,1648	0,596	0,5963	0,59601	0,59599	0,59601	0	0,596
	1,23584	2,02097	1,67712	1,91093	2,41537	1,71407	-1,1985	0,148	0,94843	-2,3452	0	-0,1171
	3,80931	4,59363	4,24982	4,48362	3,06174	4,28675	1,37492	2,7207	3,52109	0,22747	0	2,45562
ls,y	1,35329	2,13802	1,79419	2,02799	1,56927	1,83113	-1,0811	0,26507	1,06548	-2,2282	-1,1693	0
	0,5963	0,59599	0,596	0,596	0,596	0,16468	0,5963	0,596	0,59599	0,596	0,596	0
	0,06657	0,85169	0,50785	0,74165	0,28293	1,50818	-2,3678	-1,0213	-0,2208	-3,5145	-2,4556	0
	2,64001	3,42434	3,08052	3,31433	2,85561	2,15408	0,20563	1,55141	2,3518	-0,9418	0,11706	0

Stufe							Kleinste-Quadrate-Mittelwert
ls,o	A						6,6723816
ls,u	A	B					5,6135063
ls,a	A	B					5,5253194
ls,y		B	C				4,4442250
ls,e			C	D			4,1791579
ls,i			C	D	E		3,3787438
hd,a				D	E	F	3,0909347
hd,u					E	F	2,8749520
hd,i						F	2,6500394
hd,y					E	F	2,6130952
hd,o					E	F	2,4162340
hd,e					E	F	2,3062082

Figure F.1: **Synthetic voice.** Top: Student t for each target vowel. Bottom: Resulting subgroups of the target vowels (and least square mean for each vowel). Levels that are not connected by the same letter are significantly different. Intended accents: hd = High German, ls = Saxon. Vowel identifiers: a /øː/, e /eː/, i /iː/, o /oː/, u /uː/, y /yː/.



α= 0,050

KQMittelwert [j]

KQMittelwert [i]

Mittelwert[i]-Mittelwert[j] Std.-Fehlerdiff. Diff. KI unten Diff. KI oben	hd,a	hd,e	hd,i	hd,o	hd,u	hd,y	ls,a	ls,e	ls,i	ls,o	ls,u	ls,y
hd,a	0 0 0 0	-0,3257 0,54811 -1,4904 0,83894	-0,7468 0,5481 -1,9115 0,41783	-0,0815 0,54812 -1,2462 1,08313	-0,0128 0,54812 -1,1774 1,15192	-0,2711 0,54805 -1,4357 0,89344	-3,356 0,25582 -3,8583 -2,8537	-1,742 0,5481 -2,9066 -0,5773	-2,3989 0,54811 -3,5635 -1,2342	-4,6108 0,54812 -5,7754 -3,4461	-4,257 0,54812 -5,4216 -3,0923	-3,9892 0,54805 -5,1538 -2,8246
hd,e	0,32572 0,54811 -0,8389 1,49039	0 0 0 0	-0,4211 0,54816 -1,5858 0,74364	0,24418 0,54818 -0,9206 1,40895	0,31296 0,54818 -0,8518 1,47774	0,0546 0,54811 -1,1101 1,21926	-3,0303 0,54811 -4,1949 -1,8656	-1,4163 0,25628 -1,9195 -0,9131	-2,0731 0,54817 -3,2379 -0,9084	-4,285 0,54818 -5,4498 -3,1203	-3,9312 0,54818 -5,096 -2,7665	-3,6635 0,54811 -4,8281 -2,4988
hd,i	0,74682 0,5481 -0,4178 1,91147	0,4211 0,54816 -0,7436 1,58584	0 0 0 0	0,66528 0,54817 -0,4995 1,83003	0,73406 0,54817 -0,4307 1,89882	0,47569 0,5481 -0,689 1,64034	-2,6092 0,5481 -3,7738 -1,4445	-0,9952 0,54815 -2,1599 0,16955	-1,652 0,25628 -2,1552 -1,1488	-3,8639 0,54817 -5,0287 -2,6992	-3,5101 0,54817 -4,6749 -2,3454	-3,2424 0,5481 -4,407 -2,0777
hd,o	0,08155 0,54812 -1,0831 1,24623	-0,2442 0,54818 -1,4089 0,92059	-0,6653 0,54817 -1,83 0,49948	0 0 0 0	0,06879 0,54819 -1,096 1,23358	-0,1896 0,54812 -1,3543 0,9751	-3,2744 0,54812 -4,4391 -2,1098	-1,6604 0,54817 -2,8252 -0,4957	-2,3173 0,54818 -3,4821 -1,1525	-4,5292 0,25642 -5,0327 -4,0257	-4,1754 0,54819 -5,3402 -3,0106	-3,9076 0,54812 -5,0723 -2,743
hd,u	0,01276 0,54812 -1,1519 1,17744	-0,313 0,54818 -1,4777 0,85181	-0,7341 0,54817 -1,8988 0,43069	-0,0688 0,54819 -1,2336 1,09601	0 0 0 0	-0,2584 0,54812 -1,4231 0,90631	-3,3432 0,54812 -4,5079 -2,1785	-1,7292 0,54817 -2,894 -0,5645	-2,3861 0,54818 -3,5509 -1,2213	-4,598 0,54819 -5,7628 -3,4332	-4,2442 0,25642 -4,7477 -3,7407	-3,9764 0,54812 -5,1411 -2,8117
hd,y	0,27113 0,54805 -0,8934 1,4357	-0,0546 0,54811 -1,2193 1,11007	-0,4757 0,5481 -1,6403 0,68895	0,18958 0,54812 -0,9751 1,35426	0,25837 0,54812 -0,9063 1,42305	0 0 0 0	-3,0849 0,54805 -4,2494 -1,9203	-1,4709 0,5481 -2,6355 -0,3062	-2,1277 0,54811 -3,2924 -0,9631	-4,3396 0,54812 -5,5043 -3,1749	-3,9858 0,54812 -5,1505 -2,8211	-3,7181 0,25582 -4,2204 -3,2158
ls,a	3,35599 0,25582 2,8537 3,85829	3,03027 0,54811 1,8656 4,19493	2,60917 0,5481 1,44452 3,77382	3,27444 0,54812 2,10976 4,43913	3,34323 0,54812 2,17855 4,50791	3,08486 0,54805 1,92029 4,24944	0 0 0 0	1,61399 0,5481 0,44935 2,77864	0,95712 0,54811 -0,2075 2,12179	-1,2548 0,54812 -2,4194 -0,0901	-0,901 0,54812 -2,0657 0,26371	-0,6332 0,54805 -1,7978 0,53138
ls,e	1,742 0,5481 0,57735 2,90665	1,41627 0,25628 0,91308 1,91947	0,99517 0,54815 -0,1696 2,1599	1,66045 0,54817 0,49569 2,82521	1,72923 0,54817 0,56448 2,89399	1,47087 0,5481 0,30622 2,63552	-1,614 0,5481 -2,7786 -0,4493	0 0 0 0	-0,6569 0,54816 -1,8216 0,50787	-2,8688 0,54817 -4,0335 -1,704	-2,515 0,54817 -3,6797 -1,3502	-2,2472 0,5481 -3,4118 -1,0825
ls,i	2,39887 0,54811 1,2342 3,56353	2,07314 0,54817 0,90839 3,23789	1,65204 0,25628 1,14885 2,15524	2,31732 0,54818 1,15255 3,48209	2,38611 0,54818 1,22133 3,55088	2,12774 0,54811 0,96307 3,2924	-0,9571 0,54811 -2,1218 0,20754	0,65687 0,54816 -0,5079 1,82161	0 0 0 0	-2,2119 0,54818 -3,3767 -1,0471	-1,8581 0,54818 -3,0229 -0,6933	-1,5903 0,54811 -2,755 -0,4257
ls,o	4,61075 0,54812 3,44607 5,77544	4,28503 0,54818 3,12026 5,4498	3,86393 0,54817 2,69918 5,02869	4,52921 0,25642 4,02573 5,03268	4,59799 0,54819 3,4332 5,76278	4,33963 0,54812 3,17494 5,50431	1,25476 0,54812 0,09008 2,41945	2,86876 0,54817 1,704 4,03352	2,21189 0,54818 1,04712 3,37666	0 0 0 0	0,35379 0,54819 -0,811 1,51859	0,62157 0,54812 -0,5431 1,78625
ls,u	4,25696 0,54812 3,09228 5,42164	3,93124 0,54818 2,76647 5,09601	3,51014 0,54817 2,34538 4,6749	4,17541 0,54819 3,01062 5,3402	4,2442 0,25642 3,74072 4,74768	3,98583 0,54812 2,82115 5,15051	0,90097 0,54812 -0,2637 2,06565	2,51496 0,54817 1,35021 3,67972	1,85809 0,54818 0,69332 3,02286	-0,3538 0,54819 -1,5186 0,811	0 0 0 0	0,26777 0,54812 -0,8969 1,43245
ls,y	3,98919 0,54805 2,82461 5,15376	3,66346 0,54811 2,4988 4,82813	3,24237 0,5481 2,07772 4,40701	3,90764 0,54812 2,74296 5,07232	3,97643 0,54812 2,81175 5,14111	3,71806 0,25582 3,21576 4,22035	0,6332 0,54805 -0,5314 1,79777	2,24719 0,5481 1,08254 3,41184	1,59032 0,54811 0,42566 2,75499	-0,6216 0,54812 -1,7862 0,54312	-0,2678 0,54812 -1,4325 0,89691	0 0 0 0

Stufe							Kleinste-Quadrate-Mittelwert
ls,o	A						6,6211610
ls,u	A	B					6,2673666
ls,y	A	B					5,9995947
ls,a		B	C				5,3663965
ls,i			C	D			4,4092724
ls,e				D	E		3,7524017
hd,i					E	F	2,7572294
hd,e						F	2,3361308
hd,y						F	2,2815351
hd,o						F	2,0919537
hd,u						F	2,0231671
hd,a						F	2,0104060

Figure F.2: **Human voice.** Top: Student t for each target vowel. Bottom: Resulting subgroups of the target vowels (and least square mean for each vowel). Levels that are not connected by the same letter are significantly different. Intended accents: hd = High German, ls = Saxon. Vowel identifiers: a /øː/, e /eː/, i /iː/, o /oː/, u /uː/, y /yː/.

Table F.11: Lexeme contrasts within each lexeme group, by intended accent and voice. Significant contrasts (p<0.05) are marked in bold face. See also Figures 11.10 and 11.11.

Vowel	Contrasted Words	Synthetic voice		Human voice	
		Standard High German	Saxon	Standard High German	Saxon
/øː/	bloeken–loesen	t(1)= -1.867, p=0.0631	t(1)= -1.178, p=0.24	t(1)= -0.966, p=0.3368	**t(1)=2.4576, p=0.0161**
	bloeken–schoenen	**t(1)=3.915, p=0.0001**	t(1)= -0.032, p=0.9748	t(1)=1.1288, p=0.2623	**t(1)=4.3, p<0.0001**
	loesen–schoenen	**t(1)=5.6533, p<0.0001**	t(1)=1.1851, p=0.2371	**t(1)=2.09, p=0.0397**	t(1)=1.9429, p=0.0555
/eː/	leben–lehnen	t(1)=0.1924, p=0.8476	**t(1)=2.8518, p=0.0047**	t(1)=1.6701, p=0.0988	t(1)= -0.6, p=0.5499
	leben–nehmen	**t(1)= -2.952, p=0.0035**	**t(1)= -2.241, p=0.0259**	**t(1)=2.5463, p=0.0128**	t(1)=0.5879, p=0.5582
	lehnen–nehmen	**t(1)= -3.308, p=0.0011**	**t(1)= -4.753, p<0.0001**	t(1)=0.8352, p=0.4059	t(1)=1.2199, p=0.2261
/iː/	bieten–schienen	**t(1)=2.3008, p=0.0223**	t(1)=1.4205, p=0.1568	**t(1)=3.0267, p=0.0033**	**t(1)=5.1775, p<0.0001**
	bieten–sieden	**t(1)=5.361, p<0.0001**	**t(1)=4.245, p<0.0001**	t(1)=0.279, p=0.7809	t(1)=1.822, p=0.0721
	schienen–sieden	**t(1)=2.7998, p=0.0055**	**t(1)=2.9493, p=0.0035**	**t(1)= -2.691, p=0.0086**	**t(1)= -3.59, p=0.0006**
/oː/	loben–tosen	**t(1)= -6.974, p<0.0001**	t(1)=0.3455, p=0.73	**t(1)=1.8768, p=0.0642**	t(1)=1.227, p=0.2233
	loben–wohnen	t(1)= -0.347, p=0.7289	t(1)= -0.518, p=0.6047	t(1)= -0.702, p=0.4848	t(1)=1.8455, p=0.0687
	tosen–wohnen	**t(1)=6.8704, p<0.0001**	t(1)= -0.847, p=0.398	**t(1)= -2.526, p=0.0135**	t(1)=0.5554, p=0.5802
/uː/	hupen–tuten	t(1)= -1.243, p=0.2149	t(1)= -0.017, p=0.9867	t(1)=0.2129, p=0.8319	t(1)= -1.93, p=0.0572
	hupen–zoomen	**t(1)=4.7848, p<0.0001**	**t(1)=5.6563, p<0.0001**	**t(1)=2.5792, p=0.0117**	**t(1)= -3.348, p=0.0013**
	tuten–zoomen	**t(1)=5.9217, p<0.0001**	**t(1)=5.6405, p<0.0001**	**t(1)=2.5016, p=0.0144**	t(1)= -1.243, p=0.2173
/yː/	fuegen–gruessen	**t(1)= -4.245, p<0.0001**	**t(1)= -3.49, p=0.0006**	t(1)=1.1353, p=0.2597	t(1)= -1.353, p=0.1799
	fuegen–suehnen	t(1)= -0.463, p=0.6437	t(1)=1.494, p=0.1364	t(1)=1.8185, p=0.0728	t(1)=1.2295, p=0.2225
	gruessen–suehnen	**t(1)=3.7962, p=0.0002**	**t(1)=5.1449, p<0.0001**	t(1)=0.6367, p=0.5261	**t(1)=2.6577, p=0.0095**

Bibliography

Antonanzas-Barroso, Norma, Bruce R. Gerratt and Jody Kreiman (2005). Synthesizer software for modeling voice quality. *Journal of the Acoustical Society of America*, 117(4): 2544.

Atal, B. S., J. J. Chang, M. V. Mathews and J. W. Tukey (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63(5): 1535–1555.

Baayen, R. H., R. Piepenbrock and L. Gulikers (1995). *The CELEX Lexical Database (CD-ROM). Second Release*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Bachorowski, Jo-Anne, Moria J. Smoski and Michael J. Owren (2001). The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110(3): 1581–1597.

Badin, Pierre, Gérard Bailly, Lionel Revéret, Monica Baciu, Christoph Segebarth and Christophe Savriaux (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30: 533–553.

Badin, Pierre and Gunnar Fant (1984). Notes on vocal tract computation. *Quarterly Progress and Status Report*, 25(2-3): 53–108. Dept. for Speech, Music and Hearing, KTH Royal Institute of Technology.

Bailly, Lucie, Nathalie Henrich and Xavier Pelorson (2010). Vocal fold and ventricular fold vibration in period-doubling phonation: Physiological description and aerodynamic modeling. *Journal of the Acoustical Society of America*, 127(5): 3212–3222.

Barry, William J., C. E. Hoequist and F. J. Nolan (1989). An approach to the problem of regional accent in automatic speech recognition. *Computer Speech & Language*, 3(4): 355–366.

Barry, William J., Claus Nielsen and Ove Andersen (2001). Must diphone synthesis be so unnatural? *Proceedings Interspeech*, Aalborg, Denmark, 975–978.

Bell, Alexander Melville (1867). *Visible Speech: The Science of Universal Alphabetics, or Self-Interpreting Physiological Letters, for the Writing of all Languages in One Alphabet (Inaugural Edition)*. London: Simpkin, Marshall & Co. Publically available at http://www.archive.org/details/visiblespeechsc00bellgoog (accessed Feb. 14, 2014).

Benoît, Christian, Martine Grice and Valérie Hazan (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4): 381 – 392.

Benzmüller, Ralf and William J. Barry (1996a). Microsegment synthesis – Economic principles in a low-cost solution. *Proceedings 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, volume 4.

Benzmüller, Ralf and William J. Barry (1996b). Mikrosegmentsynthese – Ökonomische Prinzipien bei der Konkatenation subphonemischer Spracheinheiten. *7. Konferenz Elektronische Sprachsignalverarbeitung (ESSV1996)*, Berlin, Germany, 86–93.

Bickley, Corine and Sheri Hunnicutt (1992). Acoustic analysis of laughter. *Proceedings International Conference on Spoken Language Processing (ICSLP 1992)*, Banff, Alberta, Canada, volume 2, 927–930.

Bickley, Corine A., Kenneth N. Stevens and David R. Williams (1997). A framework for synthesis of segments based on pseudoarticulatory parameters. Jan P. H. Van Santen, Richard Sproat, Joseph P. Olive and Julia Hirschberg (Editors), *Progress in Speech Synthesis*, New York: Springer, 211–220.

Birkholz, Peter (2006). *3D-Artikulatorische Sprachsynthese*. Berlin: Logos.

Birkholz, Peter (2007a). Articulatory synthesis of singing. *Proceedings Interspeech Special Session Synthesis of Singing*, Antwerp, Belgium.

Birkholz, Peter (2007b). Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. *Proceedings Interspeech*, Antwerp, Belgium, 2865–2868.

Birkholz, Peter (2013a). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4), doi:10.1371/journal.pone.0060603. (accessed Feb. 14, 2014).

Birkholz, Peter (2013b). VocalTractLab 2.1 for Windows (Software). VTL 2.1 publically available at http://vocaltractlab.de/download-vocaltractlab/VTL2.1.zip.

Birkholz, Peter (2013c). *VocalTractLab 2.1 User Manual*. Online publication, September 24, 2013. http://vocaltractlab.de/download-vocaltractlab/VTL2.1-Manual.pdf (accessed Feb. 14, 2014).

Birkholz, Peter, Dietmar Jackel and Bernd J. Kröger (2006). Construction and control of a three-dimensional vocal tract model. *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 873–876.

Birkholz, Peter and Dietmar Jackèl (2004). Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. *Proceedings Interspeech 2004 (ICSLP)*, Jeju, Korea, 1125–1128.

Birkholz, Peter, Dietmar Jackèl and Bernd J. Kröger (2007a). Simulation of losses due to turbulence in the time-varying vocal system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4): 1218–1226.

Birkholz, Peter and Bernd J. Kröger (2006). Vocal tract model adaptation using magnetic resonance imaging. *Proceedings 7th International Seminar on Speech Production (ISSP)*, Ubatuba, Brazil, 493–500.

Birkholz, Peter and Bernd J. Kröger (2007). Simulation of vocal tract growth for articulatory speech synthesis. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 377–380.

Birkholz, Peter, Bernd J. Kröger and Christiane Neuschaefer-Rube (2011a). Model-based reproduction of articulatory trajectories for consonant-vowel-sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5): 1422–1433.

Birkholz, Peter, Bernd J. Kröger and Christiane Neuschaefer-Rube (2011b). Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. *Proceedings Interspeech 2011*, Florence, Italy, 2681–2684.

Birkholz, Peter and Christiane Neuschaefer-Rube (2012). A system for the comparison of glottal source models for articulatory speech synthesis. Abstract for the *8th International Conference on Voice Physiology and Biomechanics*. http://vocaltractlab.de/publications/birkholz-2012-icvpb.pdf (accessed Feb. 14, 2014).

Birkholz, Peter, Ingmar Steiner and Stefan Breuer (2007b). Control concepts for articulatory speech synthesis. *Proceedings 6th International Speech Communication Association (ISCA) Workshop on Speech Synthesis*, Bonn, Germany, 5–10.

Bocklet, Tobias, Andreas Maier, Josef G. Bauer, Felix Burkhardt and Elmar Nöth (2008). Age and gender recognition for telephone applications based on GMM supervectors and Support Vector Machines. *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 1605–1608.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10): 341–345. Software publically available at http://www.praat.org/ (accessed Feb. 14, 2014).

Böhtlingk, Otto (1887). *Pâṇini's Grammatik*. Leipzig: Haessel. Digitized version, publically available at http://archive.org/details/grammatikhrsgb00paniuoft (accessed Feb. 14, 2014).

Borden, Gloria J. and Katherine S. Harris (1984). *Speech Science Primer. Physiology, Acoustics, and Perception of Speech. 2nd Edition.* Baltimore, MD: Williams & Wilkins.

Bosnjak, Michael (2003). Web-basierte Fragebogenuntersuchungen – Methodische Möglichkeiten, aktuelle Themen und Erweiterungen. *Online-Erhebungen: 5. Wissenschaftliche Tagung*, Informationszentrum Sozialwissenschaften, 109–133.

Brackhane, Fabian and Jürgen Trouvain (2011). Kommentierte und typografisch bearbeitete Ausgabe der V. Abtheilung: Von der Sprachmaschine aus Wolfgangs von Kempelen k. k. wirklichen Hofraths Mechanismus der menschlichen Sprache, nebst der Beschreibung seiner sprechenden Maschine. William J. Barry and Jürgen Trouvain (Editors), *PHONUS 16. In memoriam Wolfgang von Kempelen*, Saarbrücken: Institute of Phonetics, Saarland University, 1–47.

Braun, Angelika (1988). *Zum Merkmal "Fortis/Lenis". Phonologische Betrachtungen und instrumentalphonetische Untersuchungen an einem mittelhessischen Dialekt.* Zeitschrift für Dialektologie und Linguistik. Beihefte 55, Stuttgart: Franz Steiner Verlag.

Braun, Angelika (1996). Zur regionalen Distribution von VOT im Deutschen. Angelika Braun (Editor), *Untersuchungen zu Stimme und Sprache. Papers on Voice and Speech*, Stuttgart: Franz Steiner Verlag, Zeitschrift für Dialektologie und Linguistik. Beihefte 96, 19–32.

Braun, Angelika (2001). Sprechstimmlage und regionale Umgangssprache. Angelika Braun (Editor), *Beiträge zur Linguistik und Phonetik*, Stuttgart: Franz Steiner Verlag, Zeitschrift für Dialektologie und Linguistik. Beihefte 118, 453–463.

Braun, Angelika and A. Wagner (2002). Is voice quality language-dependent? A. Braun and H. R. Masthoff (Editors), *Phonetics and Its Applications. Festschrift for Jens-Peter Köster on the Occasion of his 60th Birthday*, Stuttgart: Franz Steiner Verlag, Zeitschrift für Dialektologie und Linguistik. Beihefte 121, 298–312.

Browman, Catherine P. and Louis Goldstein (1989). Articulatory gestures as phonological units. *Phonology*, 6: 201–251, doi:10.1017/S0952675700001019. (accessed Feb. 14, 2014).

Browman, Catherine P. and Louis Goldstein (1990). Tiers in articulatory phonology, with some implications for casual speech. John Kingston and Mary E. Beckman (Editors), *Papers in Laboratory Phonology: Volume 1, Between the Grammar and Physics of Speech*, Cambridge University Press, 341–376.

Browman, Catherine P. and Louis Goldstein (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4): 155–180.

Brücke, Ernst (1876). *Grundzüge der Physiologie und Systematik der Sprachlaute für Linguisten und Taubstummenlehrer. Zweite Auflage*. Wien: Carl Gerold's Sohn. Publically available at http://www.archive.org/details/grundzgederphys01brgoog (accessed Feb. 14, 2014).

Bunnell, H. Timothy and Jason Lilley (2007). Analysis methods for assessing TTS intelligibility. *Proceedings Speech Synthesis Workshop 6 (SSW6)*, Bonn, Germany, 374–379.

Buuren, L. van (1995). Postura: clear and dark consonants, etcetera. Jack Windsor Lewis (Editor), *Studies in General and English Phonetics: Essays in Honour of Professor J.D. O'Connor*, Routledge, 130–144.

Campbell, Nick (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech and Language Processing*, 14: 1171–1178.

Campbell, Nick (2007a). Changes in voice quality due to social conditions. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 2093–2096.

Campbell, Nick (2007b). Whom we laugh with affects how we laugh. *Proceedings Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, 60–65.

Carlson, Rolf (1995). Models of speech synthesis. *Proceedings of the National Academy of Science*, volume 92, 9932–9937.

Carlson, Rolf and Björn Granström (1997). Speech synthesis. William J. Hardcastle and John Laver (Editors), *The Handbook of Phonetic Sciences*, Blackwell, 768–788.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1): 41–51.

Chafe, Wallace (2007). *The Importance of Not Being Earnest. The Feeling Behind Laughter and humor*. Amsterdam: Benjamins.

Citardi, Martin J., Eiji Yanagisawa and J. Estill (1996). Videoendoscopic analysis of laryngeal function during laughter. *The Annals of Otology, Rhinology, and Laryngology*, 105: 545–549.

Claßen, Kathrin, Grzegorz Dogil, Michael Jessen, Krzysztof Marasek and Wolfgang Wokurek (1998). Stimmqualiät und Wortbetonung im Deutschen. *Linguistische Berichte*, 174: 202–246.

Conrad, B. and P. Schönle (1979). Speech and respiration. *European Archives of Psychiatry and Clinical Neuroscience*, 226: 251–268.

Crystal, David (1997). *The Cambridge Encyclopedia of Language. Second Edition*. Cambridge University Press.

Dang, Jianwu and Kiyoshi Honda (2004). Construction and control of a physiological articulatory model. *Journal of the Acoustical Society of America*, 115(2): 853–870.

Doddington, G., W. Ligget, A. Martin, M. Przybocki and D. Reynolds (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proceedings 5th International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, 1351–1354.

Drahota, Amy, Alan Costall and Vasudevi Reddy (2008). The vocal communication of different kinds of smile. *Speech Communication*, 50: 278–287.

Draxler, Christoph (2011). Percy – an HTML5 framework for media rich web experiments on mobile devices. *Proceedings Interspeech*, Florence, Italy, 3339–3340.

Eckert, Hartwig and John Laver (1994). *Menschen und ihre Stimmen. Aspekte der vokalen Kommunikation*. Beltz PsychologieVerlagsUnion.

Edmondson, Jerold A., John Esling, Jimmy G. Harris, Li Shaoni and Lama Ziwo (2001). The aryepiglottic folds and voice quality in the Yi and Bai languages: Laryngoscopic case studies. *Mon Khmer Studies*, 31: 83–100.

Eklund, Robert (2002). Ingressive speech as an indication that humans are talking to humans (and not to machines). *Proceedings International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, USA, 837–840.

Eklund, Robert (2007). Pulmonic ingressive speech: a neglected universal? *Proceedings Fonetik 2007*, Stockholm, Sweden, 21–24.

Ekman, Paul (1977). Biological and cultural contributions to body and facial movement. John Blacking (Editor), *The Anthropology of the Body*, London: Adademic Press, 39–84.

Ekman, Paul and Wallace V. Friesen (1978). *The Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action*. Palo Alto, CA, USA: Consulting Psychologists Press.

Engwall, Olov (2006). Assessing magnetic resonance imaging measurements: Effects of sustenation, gravitation, and coarticulation. *Speech Production. Models, Phonetic Processes, and Techniques*, New York: Psychology Press, 301–314.

Esling, John H. (2002). Laryngoscopic analysis of tibetan chanting modes and their relationship to register in Sino-Tibetan. *Proceedings 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, USA, volume 2, 1081–1084.

Esling, John H. (2005). There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics*, 4: 13–44.

Fagel, Sascha, Jürgen Trouvain and Eva Lasarcyk (2009). Observing lip and vertical larynx movements during smiled speech (and laughter). *Interdisciplinary Workshop on Laughter and other Interactional Vocalisations in Speech.* http://www.coli.uni-saarland.de/conf/laughter-09/presentations/Fagel_et_al.pdf (accessed Feb. 14, 2014).

Fant, C. Gunnar M. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton & Co.

Feld, Michael (2011). *A Speaker Classification Framework for Non-Intrusive User Modeling: Speech-Based Personalization of In-Car Services*. Ph.D. thesis, Computer Science Institute, Saarland University, Saarbrücken, Germany.

Fellbaum, Klaus (2012). *Sprachverarbeitung und Sprachübertragung*. Berlin, Heidelberg: Springer.

Fels, S. Sidney, Florian Vogt, Kees van den Doel, John Lloyd and Oliver Guenter (2005). ArtiSynth: Towards realizing an extensible, portable 3D articulatory speech synthesizer. *International Workshop on Auditory Visual Speech Processing (AVSP)*, 119–124. ArtiSynth toolkit publically available at http://www.magic.ubc.ca/artisynth/pmwiki.php (accessed Feb. 14, 2014).

Fels, S. Sidney, Florian Vogt, Bryan Gick, Carol Jaeger and Ian Wilson (2003). User-centered design for an open-source 3-D articulatory synthesizer. *Proceedings 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 179–182.

Flege, James Emil (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76(3): 692–707.

Flege, James Emil and Robert Hammond (1982). Mimicry of non-distinctive phonetic differences between language varieties. *Studies in Second Language Acquisition*, 5(1): 1–17.

Fowler, Carol A. and Elliot Saltzman (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36(2,3): 171–195.

Fröhlich, Matthias, Dirk Michaelis, Hans Werner Strube and Eberhard Kruse (1997). Acoustic voice quality description: Case studies for different regions of the hoarseness diagram. *Advances in Quantitative Laryngoscopy, 2nd Round Table*, 143–150.

Fuchs, Susanne, Phil Hoole, Jana Brunner and Miki Inoue (2004). The trough effect. An aerodynamic phenomenon? *From Sound to Sense: 50 Years of Discoveries in Speech Communication*, Cambridge, MA, USA, C25–C30.

Fujisaki, Hiroya and K. Hirose (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of America*, 5(4): 233–241.

Fukui, Kotaro, Yuma Ishikawa, Eiji Shintaku, Keisuke Ohno, Nana Sakakibara, Atsuo Takanishi and Masaaki Honda (2008). Vocal cord model to control various voices for anthropomorphic talking robot. *Proceedings 8th International Seminar on Speech Production (ISSP)*, Strasbourg, France, 341–344.

Fukui, Kotaro, Toshihiro Kusano, Yoshikazu Mukaeda, Yuto Suzuki, Atsuo Takanishi and Masaaki Honda (2010). Speech robot mimicking human articulatory motion. *Proceedings Interspeech 2010*, Makuhari, Chiba, Japan, 1021–1024.

Gabelman, Brian, Jody Kreiman, Bruce R. Gerratt, Norma Antonanzas-Barroso and Abeer Alwan (1998). Perceptually-motivated modeling of noise in pathological voices. *Journal of the Acoustical Society of America*, 1293–1294.

Gay, T. and M. T. Turvey (1979). Effects of efferent and afferent interference on speech production: Implications for a generative theory of speech motor control. *Proceedings 9th International Congress of Phonetic Sciences, ICPhS*, Copenhagen, Denmark, volume 2, 344–350.

Gay, Thomas (1975). Some electromyographic measures of coarticulation in VCV utterances. *Haskins Laboratories Status Report on Speech Research (SR-44)*, 137–145.

Gick, Bryan, Ian Wilson, Karsten Koch and Clare Cook (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, 61: 220–233.

Goodin-Mayeda, C. Elizabeth (2011). Perceptual compensation for acoustic effects of nasal coupling by Spanish and Portuguese listeners. *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*, 75–83.

Grammer, Karl and Irenaus Eibl-Eibesfeldt (1990). The ritualisation of laughter. Walter A. Koch (Editor), *Natürlichkeit der Sprache und der Kultur. Bochumer Beiträge zur Semiotik*, Brockmeyer, Bochum, 192–214.

Guenther, Frank H., Satrajit S. Ghosh, Alfonso Nieto-Castanon and Jason A. Tourville (2006). A neural model of speech production. *Speech Production. Models, Phonetic Processes, and Techniques*, New York: Psychology Press, 27–39.

Gupta, S. K. and J. Schroeter (1993). Pitch-synchronous frame-by-frame and segment-based articulatory analysis-by-synthesis. *Journal of the Acoustical Society of America*, 94(5): 2517–2530.

Hammarberg, Björn (Editor) (2009). *Processes in Third Language Acquisition*. Edinburgh: Edinburgh University Press.

Hammarberg, Björn and Britta Hammarberg (2009). Re-setting the basis of articulation in the acquisition of new languages: A third language case study. Björn Hammarberg (Editor), *Processes in Third Language Acquisition*, Edinburgh: Edinburgh University Press, 74–85.

Hanson, H. M., R. S. McGowan, K. N. Stevens and R. E. Beaudoin (1999). Development of rules for controlling the HLsyn speech synthesizer. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, USA, volume 1, 85–88.

Hardcastle, William J. (1976). *Physiology of Speech Production: An Introduction for Speech Scientists*. Academic Press.

Harnsberger, James D., Rahul Shrivastav, W.S. Brown, Howard Rothman and Harry Hollien (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, 1(22): 58–69.

Hawkins, Sarah and Kenneth N. Stevens (1985). Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels. *Journal of the Acoustical Society of America*, 77(4): 1560–1575.

Heid, Sebastian and Sarah Hawkins (2000). An acoustical study of long domain /r/ and /l/ coarticulation. *Proceedings 5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Germany, 77–80.

Hill, David R., Leonard Manzara and Craig Schock (1995). Real-time articulatory speech-synthesis-by-rules. *Proceedings 14th Annual International Voice Technologies Applications Conference of the American Voice I/O Society (AVIOS)*, San Jose, USA, 27–44.

Honikman, Beatrice (1964). Articulatory settings. David Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott and Trim J. L. M. (Editors), *In Honour of Daniel Jones. Papers Contributed on the Occasion of his Eightieth Birthday 12 September 1961*, London: Longmans, 73–84.

Huckvale, Mark (2004). ACCDIST: A metric for comparing speakers' accents. *Proceedings Interspeech 2004*, 29–32.

Huckvale, Mark (2007a). ACCDIST: An accent similarity metric for accent recognition and diagnosis. Christian Müller (Editor), *Speaker Classification II. Lecture Notes in Computer Science*, Berlin: Springer, 258–275.

Huckvale, Mark (2007b). Hierarchical clustering of speakers into accents with the ACCDIST metric. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, 1821–1824.

IPDS (1995-1997). *The Kiel Corpus of Spontaneous Speech*. Institut für Phonetik und Digitale Sprachsignalverarbeitung (IPDS), Universität Kiel. CD-ROM.

IPDS (2006). *Video Task Scenario: Lindenstrasse – The Kiel Corpus of Spontaneous Speech*. Institut für Phonetik und Digitale Sprachsignalverarbeitung (IPDS), Universität Kiel. DVD.

Ishizaka, K. and J. L. Flanagan (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal*, 51(6): 1233–1268.

Jacobs, Dean, Jan Prins, Peter Siegel and Kenneth Wilson (1982). Monte Carlo techniques in code optimization. *Proceedings 15th Annual Workshop on Microprogramming*, Piscataway, NJ, USA, 143–148.

Kalos, Malvin H. and Paula A. Whitlock (2008). *Monte Carlo Methods. Second, Revised and Enlarged Edition*. Weinheim: Wiley-Blackwell.

Karlsson, Inger and J. Liljencrants (1996). Diverse voice qualities: Models and data. *Proceedings Fonetik 96, Swedish Phonetics Conference. Quarterly Progress and Status Report*. Dept. for Speech, Music and Hearing, KTH Royal Institute of Technology.

Keller, Frank, Subahshini Gunasekharan, Neil Mayo and Martin Corley (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1): 1–12.

King, Simon, Thomas Portele and Florian Höfer (1997). Speech synthesis using non-uniform units in the Verbmobil project. *Proceedings Eurospeech 97*, Rhodes, Greece, 569–572.

Kipper, Silke and Dietmar Todt (2003). The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior*, 27(4): 255–272.

Kirkpatrick, S., C. D. Gelatt Jr. and M. P. Vecchi (1983). Optimization by simulated annealing. *Science*, 220(4598): 671–680.

Klatt, Dennis H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3): 971–995.

Klatt, Dennis H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3): 737–793.

Klatt, Dennis H. and Laura C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2): 820–857.

Kreiman, Jody, Brian Gabelman and Bruce R. Gerratt (2003). Perception of vocal tremor. *Journal of Speech, Language and Hearing Research*, 46: 203–214.

Kreiman, Jody and Bruce R. Gerratt (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America*, 100(3): 1787–1795.

Kreiman, Jody, Bruce R. Gerratt and Mika Ito (2007). When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America*, 122(4): 2354–2364.

Kreul, E. James and Michael H. L. Hecker (1971). Descriptions of the speech of patients with cancer of the vocal folds. Part II: Judgments of age and voice quality. *Journal of the Acoustical Society of America*, 49(4): 1283–1287.

Kröger, Bernd J. (1998). *Ein phonetisches Modell der Sprachproduktion*. Number 387 in Linguistische Arbeiten, Tübingen: Niemeyer.

Kröger, Bernd J. and Peter Birkholz (2007). A gesture-based concept for speech movement control in articulatory speech synthesis. Anna Esposito, Marcos Faundez-Zanuy, Eric Keller and Maria Marinaro (Editors), *Verbal and Nonverbal Communication Behaviours*, Berlin, Heidelberg: Springer, volume 4775 of *Lecture Notes in Computer Science*, 174–189.

Kröger, Bernd J., Peter Birkholz, Jim Kannampuzha and Christiane Neuschaefer-Rube (2006). Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer. *Proceedings Interspeech 2006 (ICSLP)*, Pittsburgh, USA, 565–568.

Kröger, Bernd J., Peter Birkholz and Christiane Neuschaefer-Rube (2007). Ein neuronales Modell zur sensomotorischen Entwicklung des Sprechens – Modeling developmental aspects of sensorimotor control of speech production. *Laryngo-Rhino-Otologie*, 86: 365–370.

Lameli, Alfred (2004). *Standard und Substandard: Regionalismen im diachronen Längsschnitt*. Zeitschrift für Dialektologie und Linguistik. Beihefte 128, Stuttgart: Steiner.

Laprie, Yves and Anne Bonneau (2007). Construction of perception stimuli with copy synthesis. *Proceedings 16th ICPhS*, Saarbrücken, Germany, 2189–2192.

Lasarcyk, Eva (2007). Investigating larynx height with an articulatory speech synthesizer. *Proceedings 16th International Congress of the Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 2213–2216.

Lasarcyk, Eva (2010). Acoustics vs. articulation in articulatory speech synthesis: One vocal tract target configuration has more than one sound. *Proceedings Konferenz Elektronische Sprachsignalverarbeitung (ESSV'10)*, Berlin, Germany.

Lasarcyk, Eva and Jürgen Trouvain (2007). Imitating conversational laughter with an articulatory speech synthesizer. *Proceedings Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, 43–48.

Lasarcyk, Eva and Jürgen Trouvain (2008). Spread lips + raised larynx + higher F_0 = smiled speech? – An articulatory synthesis approach. *Proceedings 8th International Speech Production Seminar (ISSP)*, Strasbourg, France, 345–348.

Lasarcyk, Eva and Charlotte Wollermann (2010). Do prosodic cues influence uncertainty perception in articulatory speech synthesis? *Proceedings 7th International Speech Communication Association (ISCA) Workshop on Speech Synthesis (SSW7)*, Kyoto, Japan.

Lasarcyk, Eva, Charlotte Wollermann, Bernhard Schröder and Ulrich Schade (2013). On the modelling of prosodic cues in synthetic speech – What are the effects on perceived uncertainty and naturalness? *Proceedings 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)*, Marseille, France.

Laver, John (1978). The concept of articulatory settings: An historical survey. *Historiographia Linguistica*, 5(1-2): 1–14.

Laver, John (1980). *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics, London: Cambridge University Press.

Laver, John (1994). *Principles of Phonetics*. Cambridge: CUP.

Leky, Max (1917). *Grundlagen einer allgemeinen Phonetik*. Köln: Bachem.

Lenzo, K. and Alan W. Black (2000). Diphone collection and synthesis. *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

Levelt, W. J. M., A. Roelofs and A. A. Meyer (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences*, 22: 1–75.

Leydon, Ciara, Jay J. Bauer and Charles R. Larson (2003). The role of auditory feedback in sustaining vocal vibrato. *Journal of the Acoustical Society of America*, 114(3): 1575–1581.

Lindblom, Björn (1990). Explaining phonetic variation: A sketch of the H&H theory. *Speech Production and Speech Modelling*, Dordrecht: Kluwer, 403–439.

Lindblom, Björn, Harvey M. Sussman, Golnaz Modarresi and Elizabeth Burlingame (2002). The trough effect: Implications for speech motor programming. *Phonetica*, 59: 245–262.

Linville, Sue Ellen (2001). *Vocal Aging*. Singular.

Local, John (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics*, 31(3-4): 321–339.

Lowie, Wander and Sybrine Bultena (2007). Articulatory settings and the dynamics of second language speech production. *Proceedings Phonetics Teaching and Learning Conference (PTLC 2007)*, London, UK.

Luschei, Erich S., Lorraine O. Ramig, Eileen M. Finnegan, Kristen K. Baker and Marshall E. Smith (2006). Patterns of laryngeal electromyography and the activity of the respiratory system during spontaneous laughter. *Journal of Neurophysiology*, 96: 442–450.

Maeda, Shinji, Marie-Odile Berger, Olov Engwall, Yves Laprie, Petros Maragos, Blaise Potard and Jean Schoentgen (2008). *Deliverable D1.1 Technology Inventory of Audiovisual-to-Articulatory Inversion*. Online publication. http://aspi.loria.fr/Save/survey-1.pdf (accessed Feb. 14, 2014).

Magen, Harriet S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, 26: 381–400.

Mennen, Ineke, James M. Scobbie, Esther de Leeuw, Sonja Schaeffler and Felix Schaeffler (2010). Measuring language-specific phonetic settings. *Second Language Research*, 26(1): 13–41.

Merkel, Carl Ludwig (1866). *Physiologie der menschlichen Sprache (physiologische Lalektik)*. Leipzig: Wigand. Made publically available by Bayerische Staatsbibliothek at http://www.mdz-nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb10369014-2 (accessed Feb. 14, 2014).

Moisik, Scott R. (2008). *A Three-Dimensional Model of the Larynx and the Laryngeal Constrictor Mechanism: Visually Synthesizing Pharyngeal and Epiglottal Articulations Observed in Laryngoscopy*. Master's thesis, University of Victoria, Canada.

Moisik, Scott R. and John H. Esling (2007). 3D auditory-articulatory modeling of the laryngeal constrictor mechanism. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 373–376.

Moos, Anja and Jürgen Trouvain (2007). Comprehension of ultra-fast speech – blind vs. "normally hearing" persons. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 677–680.

Mowrer, Donald E. (1994). A case study of perceptual and acoustic features of an infant's first laugh utterances. *Humor*, 7(2): 139–155.

Mowrer, Donald E., Leonard L. LaPointe and James Case (1987). Analysis of five acoustic correlates of laughter. *Journal of Nonverbal Behavior*, 11(3): 191–199.

Müller, Christian (2005). *Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht. Two-layered context-sensitive speaker classification on the example of age and gender*. Ph.D. thesis, Computer Science Institute, Saarland University, Germany.

Müller, Christian (2006). Automatic recognition of speakers' age and gender on the basis of empirical studies. *Proceedings 9th International Conference on Spoken Language Processing (ICSLP 2006)*, Pittsburgh, USA.

Nam, Hosung, Louis Goldstein, Elliot Saltzman and Dani Byrd (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115(5): 2430–2430. Software publically available at http://www.haskins.yale.edu/tada_download/ (accessed Feb. 14, 2014).

Neiberg, Daniel, Gopal Ananthakrishnan and Olov Engwall (2008). The acoustic to articulation mapping: Non-linear or non-unique? *Proceedings Interspeech 2008*, Brisbane, Australia, 1485–1488.

Neppert, Joachim (1999). *Elemente einer akustischen Phonetik. 4., vollständig neu bearbeitete Auflage*. Hamburg: Buske.

Nieto-Castanon, Alfonso, Frank H. Guenther, Joseph S. Perkell and Hugh D. Curtin (2005). A modeling investigation of articulatory variability and acoustic stability during American English /r/ production. *Journal of the Acoustical Society of America*, 117: 3196–3212.

Nwokah, Evangeline Eva, Patricia Davies, Asad Islam, Hui-Chin Hsu and Alan Fogel (1993). Vocal affect in three-year-olds: A quantitative acoustic analysis of child laughter. *Journal of the Acoustical Society of America*, 94(6): 3076–3090.

Nwokah, Evangeline Eva, Hui-Chin Hsu, Patricia Davies and Alan Fogel (1999). The integration of laughter and speech in vocal communication: A dynamic systems perspective. *Journal of Speech, Language and Hearing Research*, 42: 880–894.

Ohala, John (1983). An ethological perspective on common cross-language utilization of F0 in voice. *Phonetica*, 41: 1–16.

Öhman, Sven E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41(2): 310–320.

Perkell, Joseph S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*. Cambridge: MIT Press.

Perkell, Joseph S. (1979). On the use of orosensory feedback: An interpretation of compensatory articulation experiments. *Proceedings 9th International Congress of Phonetic Sciences*, Copenhagen, Denmark, volume 2, 358–364.

Perkell, Joseph S. (1986). Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication*, 5(1): 47–68.

Picart, Benjamin, Thomas Drugman and Thierry Dutoit (2010). Analysis and synthesis of hypo and hyperarticulated speech. *Proceedings 7th Speech Synthesis Workshop (SSW 7)*, Kyoto, Japan.

Picart, Benjamin, Thomas Drugman and Thierry Dutoit (2011). Continuous control of the degree of articulation in HMM-based speech synthesis. *Proceedings Interspeech*, Florence, Italy, 1797–1800.

Picart, Benjamin, Thomas Drugman and Thierry Dutoit (2012). Assessing the Intelligibility and Quality of HMM-based Speech Synthesis with a Variable Degree of Articulation. *The Listening Talker Workshop (LISTA)*, Edinburgh, UK.

Pierrehumbert, Janet B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. Joan Bybee and Paul Hopper (Editors), *Frequency and the Emergence of Linguistic Structure*, Amsterdam: John Benjamins, Typological Studies in Language 45, 137–157.

Pittenger, Robert E., Charles F. Hockett and John J. Danehy (1960). *The First Five Minutes. A Sample of Microscopic Interview Analysis.* Ithaca, NY, USA: Paul Martineau.

Preuß, Simon, Christiane Neuschaefer-Rube and Peter Birkholz (2013). Prospects of EPG and OPG sensor fusion in pursuit of a 3D real-time representation of the oral cavity. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung ESSV 2013*, Dresden: TUDPress, 144–151.

Prom-on, Santitham, Peter Birkholz and Yi Xu (2013). Training an articulatory synthesizer with continuous acoustic data. *Proceedings Interspeech 2013*, Lyon, France, 349–353.

Provine, Robert R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology*, 95: 291–298.

Pucher, Michael, Dieter Schabus, Junichi Yamagishi, Friedrich Neubarth and Volker Strom (2010). Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication*, 52(2): 164–179.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Publically available at http://www.R-project.org (accessed Feb. 14, 2014).

Ramanarayanan, Vikra, Dani Byrd, Louis Goldstein and Shrikanth Narayanan (2010). Investigating articulatory setting – Pauses, ready position, and rest – Using real-time MRI. *Proceedings Interspeech 2010*, Makuhari, Japan, 1994–1997.

Reetz, Henning (2003). *Artikulatorische und akustische Phonetik*. Wissenschaftlicher Verlag Trier, 2nd edition.

Reips, Ulf-Dietrich (1997). Forschen im Jahr 2007: Integration von Web-Experimentieren, Online-Publizieren und Multimedia-Kommunikation. *CAW-97. Beiträge zum Workshop Cognition und Web*.

Reips, Ulf-Dietrich (2002a). Internet-based psychological experimenting: Five do's and five don'ts. *Social Science Computer Review*, 20(3): 241–249.

Reips, Ulf-Dietrich (2002b). Theory and techniques of Web experimenting. B. Batinic, U.-D. Reips and M. Bosnjak (Editors), *Online Social Sciences*, Seattle: Hogrefe & Huber, 229–250.

Reips, Ulf-Dietrich (2007). The methodology of Internet-based experiments. A. N. Joinson, K. McKenna, T. Postmes and U.-D. Reips (Editors), *The Oxford Handbook of Internet Psychology*, Oxford University Press, 373–390.

Richmond, Korin (2007). Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. *Proceedings Non-Linear Speech Processing (NOLISP)*, Paris, France, 67–70.

Richmond, Korin (2009). Preliminary inversion mapping results with a new EMA corpus. *Proceedings Interspeech 2009*, Brighton, UK, 2835–2838.

Richmond, Korin, Simon King and Paul Taylor (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language*, 17: 153–172.

Ridouane, Rachid, Susanne Fuchs and Phil Hoole (2006). Laryngeal adjustments in the production of voiceless obstruent clusters in Berber. *Speech Production. Models, Phonetic Processes, and Techniques*, New York: Psychology Press, 275–297.

Robson, Julie and Janet MackenzieBeck (1999). Hearing smiles - Perceptual, acoustic and production aspects of labial spreading. *Proceedings 14th International Congress of Phonetic Sciences (ICPhS)*, San Francisco, USA, 219–222.

Rothgänger, Hartmut, Gertrud Hauser, Aldo Carlo Cappellini and Assunta Guidotti (1998). Analysis of laughter and speech sounds in Italian and German students. *Naturwissenschaften*, 85(8): 394–402.

Rubin, Philip, Thomas Baer and Paul Mermelstein (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70(2): 321–328.

Ruch, Willibald and Paul Ekman (2001). The expressive pattern of laughter. Alfred W. Kaszniak (Editor), *Emotion, Qualia and Consciousness*, Singapore: Word Scientific Publishing, 426–443.

Rues, Beate, Beate Redecker, Evelyn Koch, Uta Wallraff and Adrian P. Simpson (2007). *Phonetische Transkription des Deutschen. Ein Arbeitsbuch. 2., überarbeitete und ergänzte Auflage*. Tübingen: Gunter Narr.

SAS Institute Inc. (1989–2013). *JMP. Version 10. (Statistics Software)*. Cary, NC, USA.

Schaeffer, Natalie, Stephen A. Cavallo, Meryl Wall and Carol Diakow (2002). Speech breathing behavior in normal and moderately to severely dysphonic subjects during connected speech. *Journal of Medical Speech-Language Pathology*, 10: 1–18.

Schaeffler, Sonja, James M. Scobbie and Ineke Mennen (2008). An evaluation of inter-speech postures for the study of language-specific articulatory settings. *Proceedings 8th International Speech Production Seminar (ISSP 2008)*, Strasbourg, France, 121–124.

Scheffers, Michel T. M. and Adrian P. Simpson (1995). LACS: Label assisted copy synthesis. *Proceedings 13th International Congress on Phonetic Sciences (ICPhS)*, Stockholm, Sweden, volume 2, 346–349.

Schötz, Susanne (2001). A perceptual study of speaker age. *Lund University, Dept. of Linguistics Working Papers*, 49: 136–139.

Schötz, Susanne (2003). Speaker age: A first step from analysis to synthesis. *Proceedings 15th International Congress on Phonetic Sciences (ICPhS)*, Barcelona, Spain.

Schötz, Susanne (2004). The role of F0 and duration in perception of female and male speaker age. *Proceedings Speech Prosody*, Nara, Japan, 379–382.

Schötz, Susanne (2006). *Perception, Analysis and Synthesis of Speaker Age*. Ph.D. thesis, Lund University.

Schötz, Susanne (2007a). Acoustic analysis of adult speaker age. Christian Müller (Editor), *Speaker Classification I, Lecture Notes in Computer Science*, Springer, 88–107.

Schötz, Susanne (2007b). Analysis and Synthesis of Speaker Age. *Proceedings 16th International Congress on Phonetic Sciences (ICPhS)*, Saarbrücken, Germany.

Schröder, Marc, Veronique Auberge and Marie-Agnes Cathiard (1998). Can we hear smile? *Proceedings International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia. Paper 0439.

Schröder, Marc and Jürgen Trouvain (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6: 365–377. MARY publically available at http://mary.dfki.de/ (accessed Feb. 14, 2014).

Schroeter, Juergen and Man Mohan Sondhi (1994). Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1): 133–150.

Shadle, Christine H. and Robert I. Damper (2001). Prospects for articulatory synthesis: A position paper. *Proceedings 4th International Speech Communication Association (ISCA) Workshop on Speech Synthesis*, Pitlochry, Scotland, UK, 121–126.

Shirai, Katsuhiko (1993). Estimation and generation of articulatory motion using neural networks. *Speech Communication*, 13(1-2): 45–51.

Shriberg, Elizabeth (2007). Higher-level features in speaker recognition. Christian Müller (Editor), *Speaker Classification I*, Heidelberg: Springer, volume 4343 of *LNAI*.

Sievers, Eduard (1901). *Grundzüge der Phonetik: Zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen. Fünfte verbesserte Auflage*. Leipzig: Breitkopf & Härtel. Publically available at http://www.archive.org/details/grundzgederphon00sievgoog (accessed Feb. 14, 2014).

Sjölander, Kåre and Jonas Beskow (2000). Wavesurfer – An open source speech tool. *Proceedings International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 464–467.

Smith, Rachel (2004). *The Role of Fine Phonetic Detail in Word Segmentation*. Ph.D. thesis, King's College, University of Cambridge.

Steiner, Ingmar (2010). *Observations on the Dynamic Control of an Articulatory Synthesizer Using Speech Production Data*. Ph.D. thesis, Universität des Saarlandes.

Stevens, Kenneth N. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17: 3–45.

Stevens, Kenneth N. (1998). *Acoustic Phonetics*. Cambridge: MIT Press.

Stone, Maureen (2010). Laboratory techniques for investigating speech articulation. William J. Hardcastle, John Laver and Fiona E. Gibbon (Editors), *The Handbook of Phonetic Sciences, 2nd Edition*, Wiley-Blackwell, 9–38.

Story, Brad H. and Ingo R. Titze (2002). A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function. *Journal of Phonetics*, 30: 485–509.

Strik, Helmer and Louis Boves (1992). Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, 20: 15–25.

Sundaram, Shiva and Shrikanth Narayanan (2007). Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1): 527–535.

Sundberg, Johan and A. Askenfelt (1981). Larynx height and voice source. A relationship? *Quarterly Progress and Status Report*, 22: 23–36. Dept. for Speech, Music and Hearing, KTH Royal Institute of Technology.

Sundberg, Johan, Björn Lindblom and Johan Liljencrants (1992). Formant frequency estimates for abruptly changing area functions: A comparison between calculations and measurements. *The Journal of the Acoustical Society of America*, 91(6): 3478–3482.

Sweet, Henry (1877). *A Handbook of Phonetics*. Clarendon Press. Publically available at http://www.archive.org/details/ahandbookphonet00sweegoog (accessed Feb. 14, 2014).

Sweet, Henry (1890a). *A Primer of Phonetics*. Oxford University Press. Publically available at http://archive.org/details/primerofphonetic014224mbp (accessed Feb. 14, 2014).

Sweet, Henry (1890b). *A Primer of Spoken English*. Clarendon Press. Publically available at http://www.archive.org/details/aprimerspokenen00sweegoog (accessed Feb. 14, 2014).

Tartter, Vivien C. (1980). Happy Talk: Perceptual and acoustic effects of smiling on speech. *Perception and Psychophysics*, 27(1): 24–27.

Tartter, Vivien C. and David Brown (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96(4): 2101–2107.

Taylor, Paul (2009). *Text-to-Speech Synthesis*. Cambridge: CUP.

Tiede, Mark K., Shinobu Masaki and Eric Vatikiotis-Bateson (2000). Contrasts in speech articulation observed in sitting and supine conditions. *Proceedings 5th Seminar on Speech Production*, Kloster Seeon, Germany, 25–28.

Titze, Ingo R. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area. *Journal of the Acoustical Society of America*, 83(2): 570–580.

Titze, Ingo R. (1994). *Principles of Voice Production*. Engelwood Cliffs: Prentice Hall.

Titze, Ingo R. (1995). Singing: A story of training entrained oscillators. *Journal of the Acoustical Society of America*, 97: 704.

Titze, Ingo R., Brad Story, Marshall Smith and Russel Long (2002). A reflex resonance model of vocal vibrato. *Journal of the Acoustical Society of America*, 111(5): 2272–2282.

Tokuda, Keiichi, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi and Tadashi Kitamura (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *Proceedings ICASSP*, 1315–1318. HTS toolkit publically available at http://hts.sp.nitech.ac.jp (accessed Feb. 14, 2014).

Trager, George L. (1958). Paralanguage: A first approximation. *Studies in Linguistics*, 13: 1–12.

Trager, George L. (1961). The typology of paralanguage. *Anthropological Linguistics*, 3(1): 17–21. http://www.jstor.org/stable/30022290 (accessed Feb. 14, 2014).

Trouvain, Jürgen and Fabian Brackhane (2010). Zur heutigen Bedeutung der Sprechmaschine von Wolfgang von Kempelen. *Band 2 der Tagungsbände der 20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV'09)*, Dresden, Germany, 97–107.

Trouvain, Jürgen and Fabian Brackhane (2011a). The Relevance Today of Wolfgang von Kempelen's 'Speaking Machine'. William J. Barry and Jürgen Trouvain (Editors), *PHONUS 16. In memoriam Wolfgang von Kempelen*, Saarbrücken: Institute of Phonetics, Saarland University, 149–166.

Trouvain, Jürgen and Fabian Brackhane (2011b). Wolfgang von Kempelen's 'Speaking Machine' as an instrument for demonstration and research. *Proceedings 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, 164–167.

Trouvain, Jürgen (2001). Phonetic aspects of “speech-laughs”. *Proceedings Conference on Orality and Gesturality (ORAGE)*, Aix-en-Provence, France, 634–639.

Trouvain, Jürgen (2003). Segmenting phonetic units in laughter. *Proceedings 15th International Conference of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 2793–2796.

Trouvain, Jürgen and Marc Schröder (2004). How (not) to add laughter to synthetic speech. *Proceedings of the Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany, 229–232.

Vettin, Julia (2003). *Laughter in Conversation: Acoustic Structure, Evaluation and its Relationship to Contextual Features*. Ph.D. thesis, Freie Universität Berlin.

Vettin, Julia and Dietmar Todt (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2): 93–115.

Vogt, Florian, Oliver Guenther, Alan Hannam, Kees van den Doel, John Lloyd, Leah Vilhan, Rahul Chander, Justin Lam, Charles Wilson, Kalev Tait, Donald Derrick, Ian Wilson, Carol Jaeger, Bryan Gick, Eric Vatikiotis-Bateson and Sidney Fels (2005). ArtiSynth – Designing a modular 3D articulatory speech synthesizer. *Journal of the Acoustical Society of America*, 117(4): 2542.

Wakita, H. and Gunnar Fant (1978). Toward a better vocal tract model. *Quarterly Progress and Status Report*, 19(1): 9–29. Dept. for Speech, Music and Hearing, KTH Royal Institute of Technology.

Walsh, Michael, Bernd Möbius, Travis Wade and Hinrich Schütze (2010). Multilevel exemplar theory. *Cognitive Science*, 34: 537–582.

Walter, Matthias (2006). *Sprachgesteuerte Gesichtsanimation*. Studienarbeit am Institut für Computergrafik, Universität Rostock.

Wängler, Hans-Heinrich (1972). *Physiologische Phonetik. Eine Einführung*. Marburg: Elwert.

Wilhelms-Tricarico, Reiner (1996). A biomechanical and physiologically-based vocal tract model and its control. *Journal of Phonetics*, 24(1): 23 – 38.

Williams, David R. (1996). Synthesis of initial (/s/-)stop-liquid clusters using HLsyn. *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, 2219–2222.

Wilson, Ian and Bryan Gick (2006). Ultrasound technology and second language acquisition research. *Proceedings 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)*, Somerville, MA, USA, 148–152.

Wilson, Ian Lewis (2006). *Articulatory Settings of French and English monolingual and Bilingual Speakers*. Ph.D. thesis, University of British Columbia.

Wollermann, Charlotte and Eva Lasarcyk (2007). Modeling and perceiving of different degrees of certainty in articulatory speech synthesis. *Proceedings 6th International Speech Communication Association (ISCA) Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 40–45.

Wollermann, Charlotte, Eva Lasarcyk, Ulrich Schade and Bernhard Schröder (2013). Disfluencies and uncertainty perception – Evidence from a human-machine scenario. *Proceedings 6th Workshop on Disfluency in Spontaneous Speech (DiSS 2013). Speech, Music and Hearing Quarterly Progress and Status Report. TMH-QPSR 54(1)*, Stockholm, Sweden, 73–76.

Wolters, Maria, Ravichander Vipperla and Steve Renals (2009). Age recognition for spoken dialogue systems: Do we need it? *Proceedings Interspeech 2009*, Brighton, UK.

Xu, Yi and S. Chuenwattanapranithi (2007). Perceiving anger and joy in speech through the size code. *Proceedings 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 2105–2108.

Xu, Yi and Q. Emily Wang (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33: 319–337.

Issues of PHONUS published between 1995 and 2014

Barry, W.J., B. Möbius & J. Trouvain (eds.) (2014). PHONUS 17: Eva Lasarcyk: *Empirical evaluation of the articulatory synthesizer VocalTractLab as a discovery tool for phonetic research: Articulatory-acoustic investigations of paralinguistic speech phenomena.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Trouvain (eds.) (2011). PHONUS 16. *In memoriam Wolfgang von Kempelen.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. (ed.) (2009). PHONUS 15: Manfred Pützer: *Die Rolle kortikaler und subkortikaler Strukturen bei der Initiierung und Produktion differenzierter CV-Silbenwiederholungen. Eine fMRT-Studie*. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Trouvain (eds.) (2008). PHONUS 14: Dominika Oliver: *Modelling Polish Intonation for Speech Synthesis.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. (ed.) (2008). PHONUS 13: Manfred Pützer: *Stimmqualität und Artikulation bei Dysarthrophonien.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. (ed.) (2007). PHONUS 12: Bistra Andreeva: *Zur Phonetik und Phonologie der Intonation der Sofioter-Varietät des Bulgarischen.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Trouvain (eds.) (2007). PHONUS 11: Roland Marti: *ó w dolnoserbšćinje (ó in Lower Sorbian / ó im Niedersorbischen).* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Trouvain (eds.) (2006). PHONUS 10: Caren Brinckmann: *Improving Prosody Prediction for Speech Synthesis – With and Without Symbolic Prosody Features.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Trouvain (eds.) (2005). PHONUS 9. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. (ed.) (2004). PHONUS 8: Jürgen Trouvain: *Tempo Variation in Speech Production. Implications for Speech Synthesis.* Saarbrücken: Institute of Phonetics, Saarland University of the Saarland.

Barry, W.J. (ed.) (2004). PHONUS 7: Marc Schröder: *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & M. Pützer (eds.) (2002). PHONUS 6. *Festschrift für Max Mangold zum 80. Geburtstag.* Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Koreman, with K. Kirchhoff (eds.) (2000). PHONUS 5. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Koreman (eds.) (1999). PHONUS 4. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Koreman (eds.) (1997). PHONUS 3. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & A. Addison (eds.) (1996). PHONUS 2. Saarbrücken: Institute of Phonetics, Saarland University.

Barry, W.J. & J. Koreman (eds.) (1995). PHONUS 1. Saarbrücken: Institute of Phonetics, Saarland University.

Electronic versions (PDFs) are available online at:
http://www.coli.uni-saarland.de > Research Groups > Phonetics > Phonus