Schnell gesprochene Sprache in der Unit-Selection-Sprachsynthese: Untersuchungen zu Korpuserstellung und -aufbereitung

Donata Moers[1,2], Petra Wagner[2], Bernd Möbius[1,3], Igor Jauk[1], Filip Müllers[1]

Kurzadresse: [1]Arbeitsbereich Sprache und Kommunikation der Universität Bonn, 53115 Bonn; [2]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld; [3]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
E-Mail: {dmo,bmo,ija,fmu}@ifk.uni-bonn.de, petra.wagner@uni-bielefeld.de
Web: http://www.sk.uni-bonn.de/

Zusammenfassung

In diesem Beitrag werden Untersuchungen zur Integration eines Korpus schnell gesprochener Sprache in das Unit-Selection-Synthesesystem BOSS [1] vorgestellt. Hierfür wurden zunächst zwei Synthesekorpora aufgenommen: eines in normalem und eines in schnellem und möglichst deutlichem Sprechtempo. Eine perzeptive Evaluation der Korpusaufnahmen zeigte, dass Stimuli, die aus schnell gesprochener Sprache generiert wurden, hinsichtlich der Verständlichkeit keinen Nachteil gegenüber Stimuli besaßen, die aus normal gesprochener Sprache generiert wurden; bezüglich der Natürlichkeit wurden sie eindeutig bevorzugt. Eine anschließende automatische Segmentierung wies nur marginale Unterschiede in der Segmentierungsgenauigkeit zwischen den beiden Korpusversionen auf. Basierend auf diesen Ergebnissen wurden für beide Korpora CART-basierte Dauervorhersagemodelle erstellt. Die Vorhersagegenauigkeit war für beide Versionen ähnlich gut.

1 Einleitung

Für viele Menschen mit starken Sehbehinderungen ist die Nutzung von Sprachsynthesesystemen Teil ihres täglichen Lebens. Dabei bevorzugen sie häufig eine extrem schnelle Sprachausgabe [2, 3]. Architekturen wie Formant- oder Diphonsynthese sind in der Lage, synthetische Sprache in hoher Sprechgeschwindigkeit zu generieren; allerdings klingt die produzierte Sprache nicht natürlich. Unit-Selection-Systeme wiederum sind in der Lage, natürlicher klingende Sprachsignale zu generieren, doch schnelle Sprache wurde in diesen Systemen bisher nicht adäquat implementiert.

Die Charakteristika schnell gesprochener Sprache unterscheiden sich von den Eigenschaften von in normalem Tempo gesprochener Sprache. Je schneller jemand spricht, desto undeutlicher werden seine Äußerungen. Dies liegt vor allem an der zunehmenden Überlappung artikulatorischer Gesten. Artikulatorische Zielstellungen, die für eine deutliche Artikulation von Bedeutung sind, werden nicht mehr erreicht [4]. Derartige Koartikulations- und Reduktionsphänomene beeinträchtigen die Verständlichkeit von Sprache [5, 6]. Deshalb sollten sie während der Erstellung eines Sprachkorpus, das als Bausteininventar für die Erzeugung synthetischer Sprache dienen soll, so weit wie möglich vermieden werden.

Zur Durchführung der hier vorgestellten Untersuchungen wurde daher eine Sprecherin ausgewählt, die in der Lage war, bei maximaler Sprechgeschwindigkeit möglichst deutlich zu artikulieren [7]. Anschließend wurden zwei bezüglich des linguistischen Inhalts identische Sprachkorpora aufgenommen, eines in normaler (ca. 4 Silben pro Sekunde) und eines in maximal schneller, möglichst deutlich artikulierter Sprechgeschwindigkeit (ca. 8 Silben pro Sekunde). Eine Auswahl von gleichen Sätzen aus beiden Versionen wurde dann auf verschiedene Sprechgeschwindigkeiten beschleunigt und perzeptiv evaluiert, um festzustellen, ob die schnell und gleichzeitig so deutlich wie möglich artikulierte Sprache einen Nachteil gegenüber stärker manipulierter, ursprünglich in normalem Tempo gesprochener Sprache haben würde.

Während der Erstellung neuer Korpora für die Unit-Selection-Sprachsynthese ist die Aufbereitung des Bausteininventars einer der zeitintensivsten Arbeitsschritte. Daher wurde in einem nächsten Schritt untersucht, ob das Verfahren der automatischen Lautsegmentierung nicht nur auf die in normalem Tempo gesprochene Sprache, sondern auch auf die schnell gesprochene Sprache anwendbar ist. Wäre dies nicht der Fall, dann wäre die Aufbereitung des schnell gesprochenen Korpus mit großem zeitlichen Aufwand verbunden. Weiterhin war es unklar, ob es sinnvoll ist, für schnell gesprochene Sprache ein eigenes Dauervorhersagemodell zu erstellen. Daher wurde sowohl für die normale, als auch für die schnell gesprochene Sprache auf Basis gleicher Parameter ein CART-basiertes Dauervorhersagemodell erstellt. Abschließend wird das schnell gesprochene Korpus unter Berücksichtigung der hier vorgestellten Ergebnisse in das Unit-Selection-Synthesesystem BOSS integriert und evaluiert werden.

2 Evaluation der Korpusaufnahmen

Um zu evaluieren, ob die schnell und so deutlich wie möglich gesprochene natürliche Sprache einen perzeptiven Nachteil gegenüber dauermanipulierter, ursprünglich in normalem Tempo gesprochener Sprache hat, wurden in einem ersten Experiment 20 zufällig ausgewählte, in normalem Tempo gesprochene Sätze mit Hilfe von TD-PSOLA linear auf die Geschwindigkeit der entsprechenden schnell gesprochenen, natürlich-sprachlichen Sätze beschleunigt (ca. 8 Silben pro Sekunde). Beide Versionen desselben Satzes wurden den Versuchspersonen jeweils paarweise präsentiert, mit der Aufforderung anzugeben, welche der beiden Versionen verständlicher war. Aufgrund der Ergebnisse von Janse [5] wurde erwartet, dass in dieser Bedingung die Stimuli, die auf normal ge-

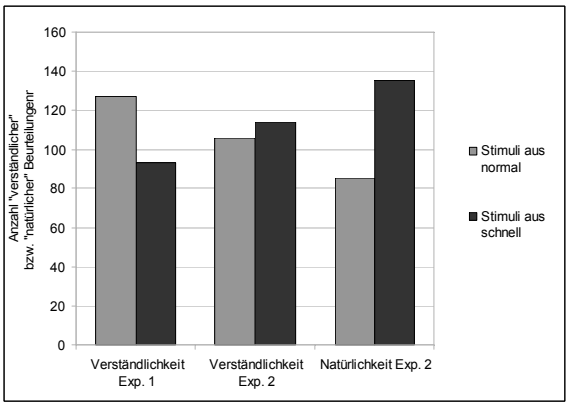

Abbildung 1: Beurteilung der Verständlichkeit von Stimuli im ersten Experiment, Beurteilung der Verständlichkeit und Natürlichkeit im zweiten Experiment.

sprochener Sprache beruhen, als besser verständlich beurteilt werden als die natürlich-sprachlichen schnell gesprochenen Stimuli.

Für ein zweites Experiment wurden anschließend beide Versionen der 20 zufällig ausgewählten Sätze mit Hilfe von TD-PSOLA linear auf ultra-schnelle (ca. 16 Silben pro Sekunde) [3] und damit unnatürlich hohe Sprechgeschwindigkeit beschleunigt. Es sollte nun sowohl die Verständlichkeit als auch die Natürlichkeit der Stimuli von den Versuchspersonen beurteilt werden. Erwartet wurde, dass die Stimuli, die aus schnell gesprochener Sprache generiert wurden, nun als etwa genauso verständlich beurteilt werden würden wie die Stimuli, die aus Sätzen in normalem Sprechtempo erzeugt wurden. Gleichzeitig würden die Stimuli, die aus schnell gesprochener Sprache generiert wurden, aufgrund der notwendigerweise stärkeren Manipulation der in normalem Tempo gesprochenen Sätze aber als wesentlich natürlicher klingend beurteilt werden.

Beide Experimente wurden direkt nacheinander in ruhiger Umgebung mit Kopfhörern durchgeführt. 11 Versuchspersonen nahmen teil. Die Ergebnisse (Abbildung 1) des ersten Experiments zeigten, dass die Stimuli, die aus normal gesprochener Sprache generiert wurden, tatsächlich als verständlicher beurteilt wurden als die natürlich-sprachlichen schnell gesprochenen Stimuli (χ^2, $p < 0.05$). Dieser Vorteil der aus normal gesprochener Sprache erzeugten Stimuli verschwand jedoch in der ultraschnellen Bedingung. Bezüglich der Verständlichkeit gab es in dieser Bedingung sogar eine leichte Tendenz hin zur Bevorzugung der Stimuli, die auf schnell gesprochener Sprache basierten, allerdings war diese nicht signifikant. Bezüglich der Natürlichkeit wurden die auf schneller Sprache basierenden Stimuli im zweiten Experiment jedoch deutlich bevorzugt (χ^2, $p < 0.01$). Diese Ergebnisse bestätigten unsere Erwartungen.

3 Automatische Segmentierung

Da die manuelle Aufbereitung eines Bausteininventars für die Unit-Selection-Sprachsynthese extrem zeitaufwändig ist, werden hierzu häufig automatische Verfahren verwendet. Die Qualität synthetischer Sprache hängt da-

Lautklasse	Anzahl Grenzen (n)	Prozent (n)	Anzahl Grenzen (s)	Prozent (s)
Lange Vokale	1641	97.3	1650	97.8
Kurze Vokale	3794	97.9	3739	96.5
Diphthonge	1256	95.1	1280	96.9
Stimmhafte Frikative	603	88.4	639	93.8
Stimmlose Frikative	1847	95.6	1887	97.6
Stimmhafte Plosive	1377	94.1	1293	88.3
Stimmlose Plosive	2973	93.8	2812*	88.7
Nasale	2403	97.0	2373	95.8
Andere (/r/, /i/, /l/)	1197*	86.4	1321	95.4

Tabelle 1: Segmentierungsgenauigkeit für einzelne Lautklassen in einem 20ms-Toleranzintervall in normaler und schneller Sprache.

bei zu einem großen Teil von der Genauigkeit der Segmentierung ab [8]. Basiert das Korpus auf schnell gesprochener Sprache, liefert ein Segmentierungsalgorithmus möglicherweise wesentlich schlechtere Ergebnisse als für normal gesprochene Sprache. Sollte dem so sein, wäre eine automatische Lautsegmentierung nicht für das schnell und deutlich gesprochenes Korpus geeignet.

Um zu evaluieren, ob die schnell gesprochene Sprache mit der gleichen Präzision wie die in normalem Tempo gesprochene Sprache automatisch segmentiert werden kann, wurde ein HTK-basierter Aligner [9] eingesetzt Gleichzeitig wurden beide Korpora manuell segmentiert. Zur Maximierung der Konsistenz der manuellen Segmentierung wurde diese Aufgabe von nur einer Person durchgeführt. Anschließend wurden die Segmentierungsfehler für alle Laute berechnet, indem die zeitliche Differenz zwischen manuell gesetzter und automatisch gesetzter Lautgrenze für jeden einzelnen Laut berechnet wurde. War die Differenz positiv, bedeutete dies, dass die automatisch erzeugte Grenze in Bezug auf die manuell gesetzte Grenze zu spät gesetzt wurde. War die Differenz negativ, dann war die automatisch erzeugte Grenze zu früh gesetzt worden.

In der Literatur finden sich Untersuchungen zur Übereinstimmung zwischen menschlichen Segmentierern, die im besten Fall bei 94% übereinstimmender Grenzen innerhalb eines 20ms-Intervalls liegt [10]. Die Auswertung der hier durchgeführten automatischen Segmentierung zeigte, dass die Segmentierungsgenauigkeit für beide Sprechgeschwindigkeiten ähnlich hoch war: In der in

normalem Tempo gesprochenen Sprache lagen 90% aller automatisch erzeugten Lautgrenzen innerhalb des 20ms-Intervalls um die manuell gesetzte Lautgrenze herum, in der schnell gesprochenen Sprache waren es 91% aller Lautgrenzen. Wird die Segmentierungsgenauigkeit jedoch bezüglich einzelner Lautklassen betrachtet (Tabelle 1), zeigen sich signifikante Unterschiede: In schnell gesprochener Sprache werden die Laute /r/, /l/ und /j/ signifikant besser segmentiert als in normal gesprochener Sprache (χ^2, $p < 0.05$). Bei stimmlosen Plosiven verhält es sich umgekehrt; sie werden in normal gesprochener Sprache signifikant besser segmentiert als in schnell gesprochener Sprache (χ^2, $p < 0.05$).

Da aber vor allem die aus automatisch segmentierter schneller Sprache synthetisierten Äußerungen Mängel in Form von fehlenden Segmenten aufwiesen, scheint trotz der guten globalen Ergebnisse eine – zumindest partielle – manuelle Korrektur der automatischen Segmentierung vonnöten zu sein.

4 Dauerprädiktion

Die Dauer phonetischer Segmente ist ein weiterer wichtiger Faktor bei der Erzeugung natürlich klingender Sprache [11]. Weil es unklar war, ob für schnell gesprochene Sprache ein valides Dauervorhersagemodell erzeugt werden kann, wurde unter Berücksichtigung der vorherigen Erkenntnisse jeweils ein CART-basiertes Dauervorhersagemodell [12] für die in normalem Tempo sowie für die schnell gesprochene Sprache generiert. Hierbei wurden wichtige phonetische Merkmale, welche die Segmentdauer beeinflussen, berücksichtigt.

Es wurde erwartet, dass die Dauervorhersage eine signifikant höhere Korrelation zwischen beobachteter und prädizierter Dauer für die normal gesprochene Sprache aufzeigt, da schnell gesprochene Sprache aufgrund vermehrt auftretender Koartikulations- und Reduktionsphänomene mehr Variabilität aufweist als in normalem Tempo gesprochene Sprache. Sollten diese Phänomene aber tatsächlich in der schnell gesprochenen Sprache weitestgehend vermieden worden sein, würde die Korrelation zwischen beobachteter und prädizierter Dauer für beide Sprechgeschwindigkeiten ähnlich hoch sein.

Für die Erstellung der CART-basierten Dauerprädiktionsmodelle wurde das Werkzeug *wagon* von den Edinburgh Speech Tools [13] benutzt. Das Merkmalsset wurde an die Anforderungen des Sprachsynthesesystems BOSS [1] angepasst und beinhaltete folgende Merkmale:

- Phonidentität
- Phondauer
- Vorhergehendes Phonem
- Nachfolgendes Phonem
- Übernächstes Phonem
- Position in der Phrase
- Betonung

Das Phon selbst ist das Merkmal, dessen Dauer vorhergesagt werden soll. Die Trainingsdaten bestanden aus den Dauern der einzelnen Phone, die aus dem jeweiligen Korpus extrahiert wurden. Die Position in der Phrase war entweder initial, medial oder final. Die Betonung konnte die Werte 1 (Hauptbetonung), 2 (Nebenbetonung) oder 0 (keine Betonung) annehmen. Ebenso wie die Dauer wurden auch der Laut selbst, der vorhergehende und nachfolgende Laut sowie die Betonung direkt aus dem Korpus extrahiert, nachdem das Korpus bereits segmentiert und durch die Nutzung der BOSS-Tools [14] vorbereitet worden war. Das übernächste Phon und die Position des Lautes in der Phrase wurden während des Prozesses berechnet.

Normale Sprechgeschwindigkeit	Schnelle Sprechgeschwindigkeit
1. Phonidentität: 0.4734	1. Phonidentität: 0.4736
2. Position in Phrase: 0.6750	2. Position in Phrase: 0.6649
3. Nachfolgendes Phonem: 0.7862	3. Nachfolgendes Phonem: 0.7559
4. Vorhergehendes Phonem: 0.8000	4. Vorhergehendes Phonem: 0.7681
5. Betonung: 0.8009	5. Betonung: 0.7738
6. Übernächstes Phonem: 0.8018	6. Übernächstes Phonem: 0.7749

Tabelle 2: Rangfolge der für die CART-basierte Dauerprädiktion verwendeten Merkmale.

Die Ergebnisse zeigten, dass die Korrelation zwischen beobachteter und vorhergesagter Dauer für die schnell gesprochene Sprache 0,78 betrug, während sie für die in normalem Tempo gesprochene Sprache 0,80 betrug. Dies ist kein signifikanter Unterschied; beide Werte entsprechen Korrelationen, die für andere Sprachen in normalem Sprechtempo berichtet werden [15]. Der mittlere quadratische Fehler sowie der durchschnittliche (absolute) Fehler waren für schnell gesprochene Sprache zwar kleiner, jedoch ist dies der kürzeren Segmentdauer in schnell gesprochener Sprache zuzuschreiben und nicht als größere Vorhersagegenauigkeit zu interpretieren.

Auch mit Blick auf die Rangfolge der verwendeten Merkmale (Tabelle 2), welche mit der Option *stepwise* von *wagon* [13] erzeugt wurde, zeigten sich keine signifikanten Unterschiede zwischen den beiden Versionen. Das wichtigste Merkmal war das Phon selber, gefolgt von seiner Position in der Phrase. Überraschenderweise weist auch der Stellenwert der Betonung keinen Unterschied zwischen normaler und schnell gesprochener Sprache auf, obwohl dies zu erwarten war, da die Anzahl und Stärke der Betonungen in schnell gesprochener Sprache in der Regel abnehmen. Dies ist möglicherweise ein Hinweis darauf, dass das Vorhaben, die schnelle Sprache mit größtmöglicher Präzision zu artikulieren, gelungen ist.

Da die CART-basierten Dauerprädiktionsmodelle nur marginale Unterschiede in der Korrelation zwischen beobachteter und vorhergesagter Dauer aufweisen, ist diese

Methode offenbar nicht nur für die Erstellung eines Dauermodells für in normalem Tempo gesprochene Sprache verwendbar, sondern ebenso für die Erstellung eines Dauermodells für schnell gesprochene Sprache zu nutzen.

5 Schlussbemerkung

Ziel des Projekts war die Einbindung eines Korpus schnell gesprochener Sprache in das Unit Selection Synthesesystem BOSS. Die perzeptive Evaluation der erstellten Korpusaufnahmen zeigte, dass Stimuli, die aus schnell gesprochener Sprache generiert wurden, in der ultra-schnellen Bedingung als ebenso verständlich und gleichzeitig als weit natürlicher klingend beurteilt wurden als Stimuli, die aus normal gesprochener Sprache generiert wurden. Die notwendige starke Dauermanipulation mag einen Teil zu diesem Ergebnis beigetragen haben, da der verwendete TD-PSOLA-Algorithmus dafür bekannt ist, bei starker Manipulation Artefakte zu produzieren [16]. Ein alternativer Ansatz wäre hier möglicherweise die Verwendung nicht-linearer Beschleunigungsalgorithmen.

Die automatische Segmentierung hat nur geringfügige Unterschiede in der Segmentierungsgenauigkeit zwischen den beiden Korpusversionen ergeben. Da die aus automatisch segmentierter Sprache synthetisierten Äußerungen Mängel in Form von fehlenden Segmenten aufwiesen, scheint zumindest eine partielle manuelle Korrektur der automatischen Segmentierung nötig zu sein. Weiterhin ist die Dauer der Einzelsegmente in schnell gesprochener Sprache im Durchschnitt kürzer, so dass sich die Frage stellt, ob ein Vergleich der Segmentierungsgenauigkeit innerhalb derselben Toleranzintervalle angemessen ist, um die Segmentierbarkeit schnell gesprochener Sprache zu beurteilen.

Die CART-basierte Dauerprädiktion lieferte für beide Sprechgeschwindigkeiten ebenfalls ähnliche Ergebnisse. Dennoch schnitt auch hier die schnelle Sprache etwas schlechter ab als die normale Sprache. Dies deutet möglicherweise darauf hin, dass die verwendeten Merkmale in den verschiedenen Sprechgeschwindigkeiten unterschiedlich stark ausgeprägt vorliegen und ihre Auswahl besser an die jeweilige Bedingung angepasst werden muss.

Der nächste Schritt wird die Erzeugung synthetisierter schnell gesprochener Sprache in verschiedenen Sprechgeschwindigkeiten sein, um die Verständlichkeit und Natürlichkeit der so erzeugten Sprache zu evaluieren. Aus diesen sowie den hier vorgestellten Untersuchungsergebnissen könnten sich in Zukunft robuste Richtlinien für die Einbindung eines Korpus schnell gesprochener Sprache in ein Unit-Selection-Synthesesystem ergeben.

Literatur

[1] E. Klabbers et al. Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proceedings Eurospeech*, Aalborg, Denmark, 2001.

[2] D. Moers, P. Wagner, and S. Breuer. Assessing the Adequate Treatment of Fast Speech in Unit Selection Speech Synthesis Systems for the Visually Impaired. In *Proceedings 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 2007.

[3] A. Moos and J. Trouvain. Comprehension of Ultra-Fast Speech – Blind vs. 'Normally Hearing' Persons. In *Proceedings ICPhS XVI*, pages 677-684, Saarbrücken, Germany, 2007.

[4] K.J. Kohler. Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In *W.J. Hardcastle and A. Marchal [Ed]: Speech Production and Speech Modelling*, pages 69-92, Dordrecht, The Netherlands, 1990.

[5] E. Janse. Production and Perception of Fast Speech. *Dissertation,* Universiteit Utrecht, The Netherlands, 2003.

[6] J.C. Krause and L.D. Braida. Investigating Alternative Forms of Clear Speech: The Effects of Speaking Rate and Speaking Mode on Intelligibility. In *Journal of the Acoustical Society of America 112*, pages 2165-2172, 2002.

[7] D. Moers and P. Wagner. Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis. In *Proceedings Interspeech*, Brighton, UK, 2009.

[8] J. Kominek, C. Bennett and A.W. Black. Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In *Proceedings Eurospeech*, Geneva, Switzerland, 2003.

[9] R. Dragon. LAM4HTK. http://www.tnt.uni-hannover.de/print/staff/dragon/index.php

[10] J. Adell and A. Bonafonte. Towards Phone Segmentation for Concatenative Speech Synthesis. In *Proceedings 5th ISCA Workshop on Speech Synthesis (SSW5),* Pittsburgh, PA, 2004.

[11] R. Carlson, B. Granström, and D. Klatt. Some Notes on the Perception of Temporal Patterns in Speech. In *B. Lindblom and S. Öhman [Ed]: Frontiers of Speech Communication Research*, London, UK: Academic Press, 1979.

[12] L. Breiman et al. Classification and Regression Trees, Belmont, USA: Wadsworth, 1984.

[13] S. King, A.W. Black, P. Taylor, R. Caley, and R. Clark. Edinburgh Speech Tools Library. System Documentation Edition 1.2, for 1.2.3 24th Jan 2003. http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

[14] S. Breuer et al. Bonn Open Synthesis System (BOSS) 3 Documentation and User Manual December 20, 2005. http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/boss/BOSS_Documentation.pdf

[15] K. Klessa, M. Szymánski, S. Breuer, and G. Demenko. Optimization of Polish Segmental Duration Prediction with CART. In *Proceedings 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 2007.

[16] S.-H. Chen, S.-J. Chen, and C.-C. Kuo. Perceptual Distortion Analysis and Quality Estimation of Prosody-Modified Speech for TD-PSOLA. In *Proceedings of the ICASSP*, Toulouse, France, 2006.