MUSS SYNTHETISCHE SPRACHE IMMER LANGWEILIG KLINGEN?

Jürgen Trouvain

Phonetik-Büro Trouvain
trouvain@phonetik-buero.de

Abstract: In einer Untersuchung mit 11 verschiedenen synthetischen Stimmen wird der Frage nachgegangen, ob synthetische Sprache dem oft geäußerten Eindruck des Gelangweilten entspricht. Ein Hörtest und eine instrumentelle Analyse prosodischer Eigenschaften zeigen, dass ein angemessenes Ausnutzen intonatorischer und temporaler Parameter dem negativen Eindruck entgegenwirken kann.

1 Einführung

Man registriert oft die Meinung, dass synthetische Sprache "langweilig" und "wenig ansprechend" klingt und dass längeres Zuhören Ermüdung bewirkt. Wirkt synthetische Sprache wirklich langweilig? Dieser Frage wird in einem Hörexperiment nachgegangen. Mit Sprachproben, die von 11 verschiedenen deutschsprachigen Synthesesystemen stammen, wird dieser oft geäußerte Eindruck überprüft. Eine instrumentelle und auditive Analyse zeigen, welche prosodische Faktoren für den negativ besetzten Höreindruck verantwortlich sind. Darüber hinaus werden weitere Aspekte von "langweilig klingender" Sprachsynthese diskutiert.

2 Hörexperiment

2.1 Material

Beim Stimulusmaterial für das Perzeptionsexperiment wurden 11 verschiedene synthetische Stimmen verwendet. Alle Stimmen wurden von 6 verschiedenen "Text-to-Speech" (TTS) - Systemen durch Web-Interfaces generiert. Alle verwendeten TTS-Systemen konkatenieren Signalbausteine, die zumeist in der Größe von Diphonen vorliegen.

Zwei unterschiedlich lange Sätze wurden "vorgelesen":

- Satz 1: 7 Wörter, 12 Silben
- Satz 2: 17 Wörter, 29 Silben

Somit standen insgesamt 22 Stimuli zur Beurteilung bereit.

2.2 Experimenteller Aufbau

Jeder Stimulus wurde von 10 Versuchspersonen, allesamt Studierende zwischen 20 und 30 Jahren, nach seinem emotionalen Gehalt beurteilt. Dazu dienten drei sieben-Punkte-Skalen, die die folgenden Dichotomien darstellen:

- "zufrieden-enttäuscht"
- "interessiert-gelangweilt"
- "erfreut-traurig"

Alle Stimuli wurden in randomisierter Reihenfolge visuell auf einem PC-Bildschirm aufgeführt. Die Versuchsperson konnte sich einen Stimulus so oft sie es wünschte durch Klicken auf das entsprechende Symbol anhören, wobei die Stimuli per Kopfhörer am PC

dargeboten wurden. Nach jedem Stimulus musste die Versuchsperson ihr Urteil in den drei Kategorien auf vorgefertigten Antwortbögen durch Ankreuzen abgeben.

2.3 Ergebnisse

Die Ergebnisse in Abbildung 1 zeigen, dass keineswegs synthetische Sprache *per se* als "langweilig" (oder "traurig" oder "enttäuscht") eingestuft wurde. Unter den Sprachproben gibt es klare Favoriten - sowohl auf der positiven als auch auf der negativen Seite.

Man sieht auch, dass *innerhalb desselben TTS-Systems* Stimmen durchaus unterschiedliche Beurteilungen erzielen können. Vergleicht man Sprecher 1 und 2 von TTS 1, so erkennt man, dass beide Stimmen über einen Punkt auseinander liegen. Eine Divergenz innerhalb desselben Systems muss aber nicht der Fall sein, wie das Beispiel von Sprecher 7 und 8 von TTS 4 zeigt.

Zwischen den Systemen sind deutliche Unterschiede zu vermerken. TTS 2 und TTS 6 liegen beispielsweise ca. 3 Punkte auseinander. Ebenso finden sich die eine Hälfte der untersuchten TTS-Systeme auf der positiven Seite der Beurteilungen, wohingegen die andere Hälfte der Synthesen auf der negativen Seite zu finden sind.

Unterschiede in den Beurteilungen für die drei Dichotomien pro Stimme sind, wenn überhaupt, nur sehr gering.

Der Faktor Satzlänge hatte keinen Einfluss auf die Beurteilungen. D.h. auch bei kürzeren Sätzen urteilten die Versuchspersonen bezüglich der gefragten Attribute nicht anders als bei längeren Sätzen, in denen mehr Audiomaterial zur Verfügung stand.

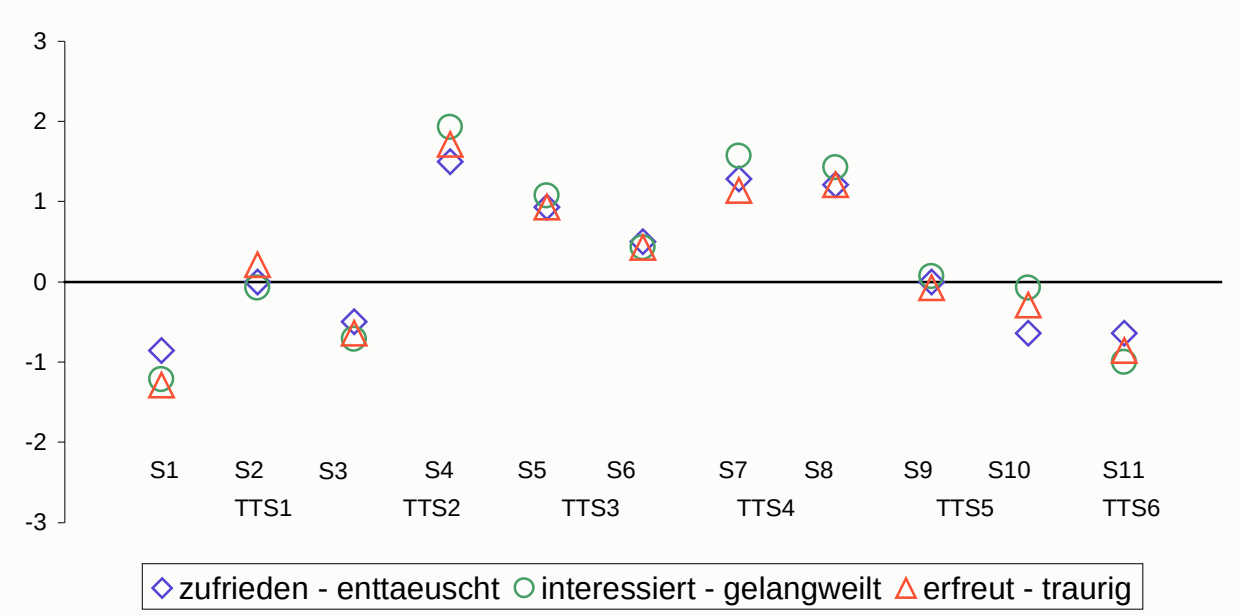

Abbildung 1 – Gemittelte Ergebnisse der Hörer-Beurteilungen auf einer 7-Punkte-Skala (von –3 bis +3) für 11 verschiedene synthetische Stimmen (S1-S11), die von 6 verschiedenen TTS Systemen stammen (TTS1-TTS6).

3 Analyse der Prosodie

Drei Beispiele am negativen bzw. positiven Ende wurden eingehender auf die prosodischen Bereiche Intonation, Rhythmus und Tempo untersucht. Für jedes Gegensatzpaar wurden die drei positivsten den drei negativsten zu einer instrumentellen Analyse gegenübergestellt.

3.1 Intonation

Bezüglich des globalen Umfangs von F0 sind die negativen Stimmen doppelt bis dreifach so wenig ausgeprägt wie die positiven Stimmen. Diese drastischen Unterschiede im Gebrauch des F0-Umfangs machen sich vor allem bei satzakzentuierten Silben und am Ende von Intonationsphrasen bemerkbar (siehe Abbildung 2). Durch den effektiveren Gebrauch des Tonhöhenumfangs steigen die positiven Stimmen steiler an bzw. fallen steiler ab. Auch ist es so, dass die negativ bewerteten im Vergleich zu den positiv bewerteten Stimmen eine geringere Anzahl an Tonhöhenbewegungen realisieren.

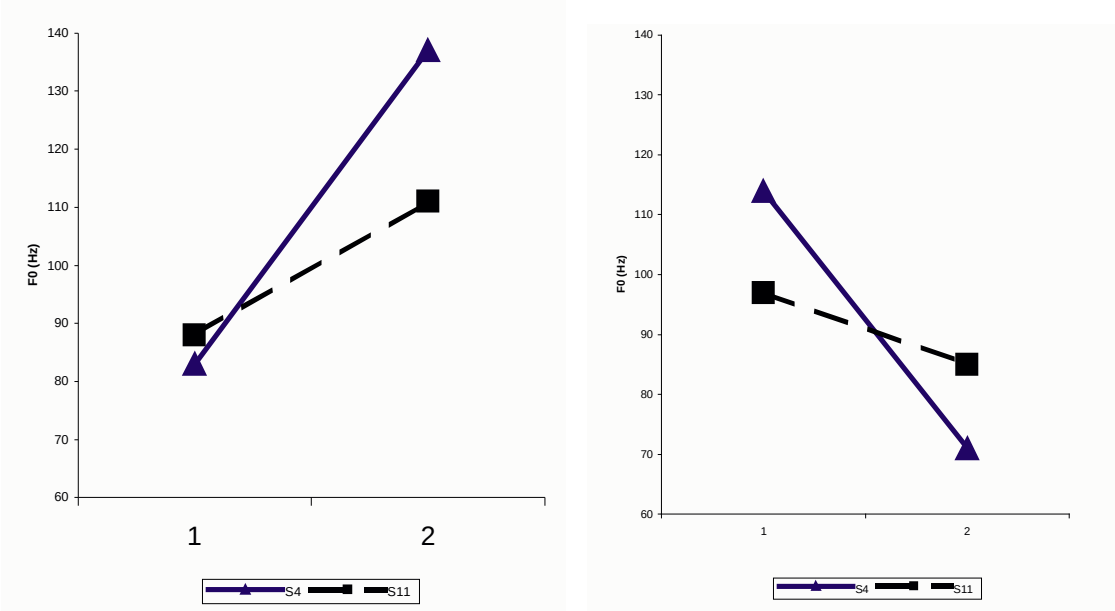

Abbildung 2 – Lokaler F0-Umfang (in Hz) an zwei markanten Stellen im ersten Satz von der Stimme, die am besten (S4) bzw. am schlechtesten (S11) beurteilt wurde: links die Tonhöhenbewegung beim ersten Satzakzent, rechts die Tonhöhenbewegung beim letzten Satzakzent zum finalen Grenzton.

3.2 Timing

In Bezug auf Eigenschaften des Timings ist es bemerkenswert, dass die als negativen eingestuften Stimmen durchweg ein langsameres Tempo als die positiver eingestuften aufwiesen (Abbildung 3). Im Durchschnitt sind dies 0,5 Silben pro Sekunde weniger. Dies trifft sowohl für die Artikulationsgeschwindigkeit (ohne Pausen) als auch für die Gesamtsprechgeschwindigkeit (inklusive Pausen) zu.

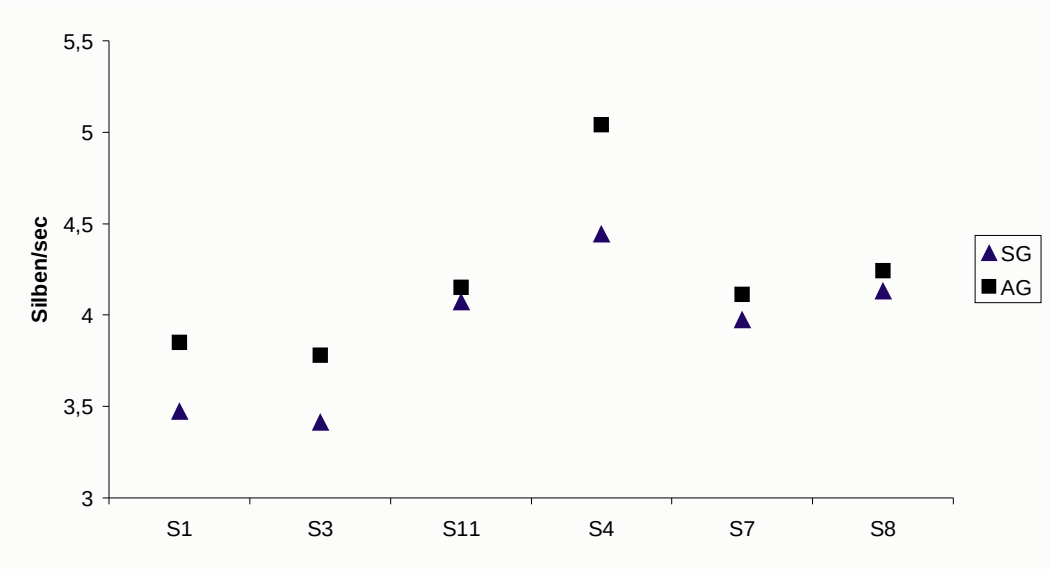

Abbildung 3 – Tempo für die drei negativsten Stimmen (S1,3,11) und die drei positivsten Stimmen (S4,7,8) in Silben pro Sekunde für die Sprechgeschwindigkeit (SG) inkl. Pausen und die Artikulationsgeschwindigkeit (AG) ohne Pausen.

Des weiteren konnte festgestellt werden, dass die negativen Stimmen durch einen kontrastärmeren Rhythmus markiert sind. In Abbildung 4 wird exemplarisch sichtbar, wie im Vergleich zu den positiven Stimmen, die akzentuierten Silben kürzer und infolgedessen die unakzentuierten Silben länger sind. Dies gilt insbesondere für den unakzentuierten Auftakt, d.h. für das Silbenmaterial vor dem ersten Satzakzent in einer Phrase. Des weiteren sind die phrasen-finale Silben ebenfalls kürzer.

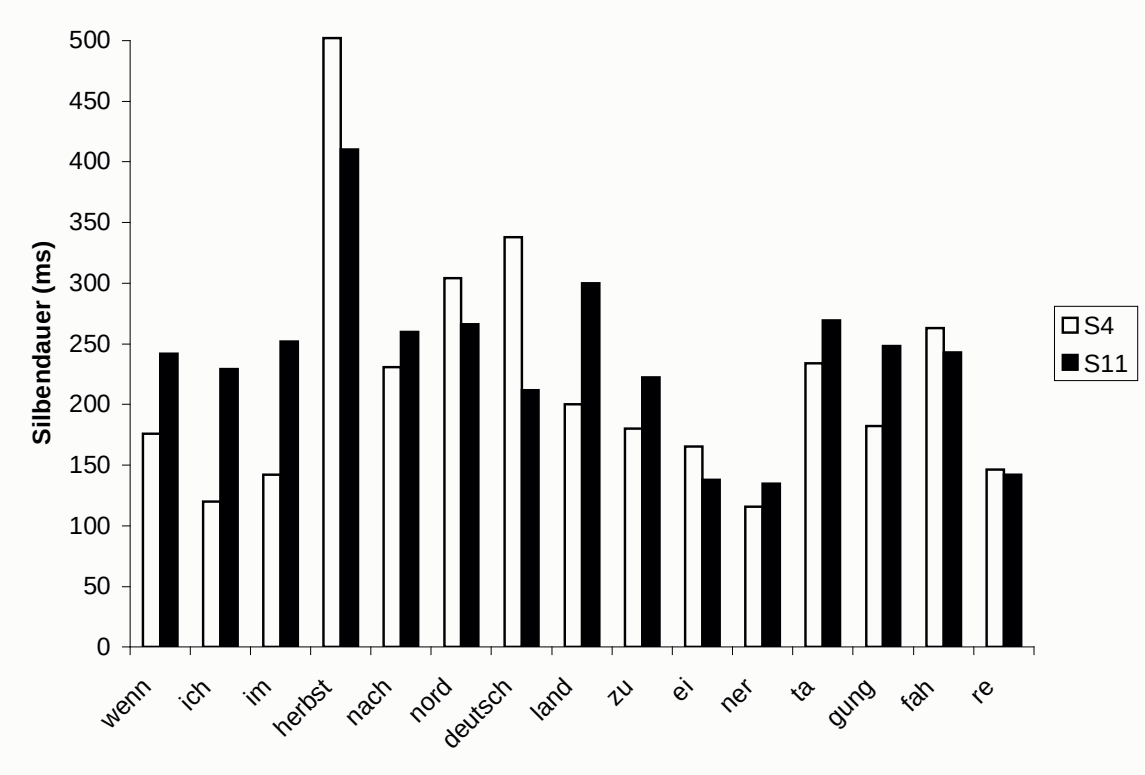

Abbildung 4 – Dauer der Silben der ersten Hälfte des langen Satzes von der Stimme, die am besten (S4) bzw. am schlechtesten (S11) beurteilt wurde.

.

3.3 Diskussion

Die im Titel gestellte Frage, ob synthetische Sprache immer langweilig klingen *muss*, kann eindeutig mit Nein beantwortet werden. Die Ergebnisse in Abbildung 1 zeigen, dass keineswegs synthetische Sprache *per se* als "langweilig" eingestuft wird. Innerhalb der untersuchten synthetischen Stimmen gibt es durchaus Beispiele, die mit "zufrieden", "interessiert" und "erfreut" in Verbindung gebracht werden. Allerdings gibt es auch solche synthetischen Stimmen, die den oft geäußerten Eindruck der Langweiligkeit synthetischer Sprache bestätigen.

Schaut man sich die prosodischen Realisierungen der einzelnen Stimmen an, so sind eindeutige Unterschiede in Bezug auf den globalen und lokalem F0-Umfang, Artikulationsgeschwindigkeit und rhythmischer Kontrastierung zwischen den "langweiligen" und "interessierten" Stimmen zu erkennen. Man kann davon ausgehen, dass allgemeine sprecherzieherische Anweisungen wie sie auch für natürliche Sprecher gelten, auch für maschinelle Sprecher gelten können: nicht zu monoton sprechen und deutlicher akzentuieren.

Es ist nicht überraschend, dass die Beurteilungen für die drei Dichotomien sehr nah beieinander liegen. Sowohl "traurig" als auch "gelangweilt" werden in der Literatur mit vermindertem durchschnittlichen F0, reduziertem F0-Umfang, langsamerem Tempo und geringerer Intensität beschrieben (vgl. [1][2][7]). Aus diesem Grund verwundert es nicht, dass "traurig", "gelangweilt" und "neutral" auch in natürlicher Sprache oft verwechselt werden.

4 Allgemeine Diskussion

4.1 neutral = nicht gelangweilt?

Die oben erwähnte Verwechslungsanfälligkeit von "neutral" und "traurig" in natürlicher Sprache sollte auch in synthetischer Sprache zu finden sein. Wie die positiven Beispiele oben gezeigt haben, ist es - bei entsprechender Modellierung der Prosodie - durchaus möglich, sich dem negativen Eindruck durch diese Verwechslungsanfälligkeit zu entziehen.

Es ist zu erwarten, dass solche Sprachsynthesen, die auf einer "Non-uniform Unit Selection" basieren, nicht mit dieser negativ auffälligen Prosodie zu tun haben und als "nicht langweilig" gelten, da sie die größtmögliche natürliche Prosodie bereits durch variabel große Signalabschnitte und eine Vielzahl von Signalabschnitten natürlicher Sprache abdecken.

4.2 Langsames Sprechtempo: Verständlichkeit vs. Langeweile

Da es Anzeichen gibt, dass viele Hörer ein Sprechtempo (inklusive Pausen) wünschen, das langsamer ist als das Tempo in natürlicher Sprache [10], kann sich das Problem des Eindrucks von Langeweile mit einer langsamen Default-Artikulationsgeschwindigkeit noch verschärfen. Entsprechende Gegenmaßnahmen könnten zum einen einen adäquate Vorhersage von Satzakzenten und Intonationskonturen sowie der Phrasierung umfassen. Wie die vorliegende prosodische Analyse gezeigt hat, scheint es zum anderen sehr wichtig zu sein, diese phonologischen Information in akustische Parameter wie Zielpunkte der Grundfrequenz sowie Silben- und Pausendauer adäquat umzusetzen. Optimalerweise würde die Modellierung durch Hörtests verifiziert werden.

4.3 Prosodische Makrostruktur

Der eingangs formulierte Vorwurf an Sprachsynthese war auch, dass längeres Zuhören Ermüdung bewirkt. Da in dieser Studie - wie in vielen anderen Studien auch - nur ein

Einzelsatz-Test-Paradigma herangezogen wurde, lässt sich keine Aussage darüber treffen, ob der Eindruck der Langeweile entsteht, wenn das zu beurteilende Audiomaterial in Größen von mehreren Minuten anstatt von mehreren Sekunden dargeboten wird.

Es kann die Vermutung angestellt werden, dass hintereinander gereihte Sätze, die als Einzelsatzäußerungen akzeptabel waren, in ihrer Gesamtheit zu einer monotonen prosodischen Makrostruktur führen. Hierbei ist zu bemerken, dass Prosodie nicht nur den supra-segmentalen Bereich abdeckt, d.h. über der Größe eines Lautsegmentes, sondern auch im Bereich oberhalb einzelner Phrasen, sozusagen "supra-phrasal" wirkt. Erste Ansätze hierzu finden sich beispielsweise in der sog. "Paragraphen-Intonation" bei Lehiste [5], die beobachtet, dass fallende finale F0-Konturen am Ende von Textabschnitten ("Paragraphen") tiefer enden als im Vergleich zum Binnentext. Aber auch die Zugehörigkeit zu verschiedenen Diskurseinheiten des Textes kann durch verschiedenartigen F0-Umfang markiert werden [6]. Hier wäre es wünschenswert, wenn bei vorliegender diskurs-bezogener Information, wie bei Concept-to-Speech, solche diskurs-bezogenen prosodischen Besonderheiten Beachtung fänden.

4.4 Langweiliger synthetischer Dialogpartner

Synthetische Sprecher werden immer häufiger in Dialogsystemen benutzt. Hierzu sind zwei Punkte anzumerken:

- Dialoge sind üblicherweise spontane Sprache (im Gegensatz zu gelesener Sprache)
- dialogisches Sprechen verhält sich prosodisch anders als monologisches Sprechen

Spontansprache ist gegenüber gelesener Sprache durch mehr Variation im Pausenverhalten, in der Artikulationsgeschwindigkeit [11], aber auch durch mehr Satzakzente und andere Typen von Satzakzenten gekennzeichnet.

Bei synthetischen Sprechern, die im Vorlesemodus für ein Dialogsystem eingesetzt werden, könnte man die Parallele zu minder gut vortragenden Schülern ziehen, die beim Einüben eines Theaterstücks ihre Sprech-Beiträge "ohne Zusammenhang" und "leblos" realisieren, weil sie vollauf damit beschäftigt sind ihre Worte auswendig vorzusagen. Dies wirkt nicht ansprechend und im schlimmsten Fall langweilig.

Wirklich dialogische Partner gehen in aller Regel auf prosodische Parameter wie die Höhe von F0 des Dialog-Partners an Anfängen und Enden von Sprecherbeiträgen ein (vgl. [3]).

4.5 Modellierung von Emotionen

Die bislang geführte Diskussion behandelte das Abwenden des Eindrucks von "langweilig", "traurig" und "enttäuscht". Eine explizite Modellierung genau solcher Emotionen kann in bestimmten Anwendungen auch wünschenswert sein: sei es, wenn ein Benutzer solche Attribute mittels Tags in einer dafür verwendbaren Markup-Sprache für Sprachsynthese wiedergeben möchte; sei es, um emotionale Färbungen mittels Sprachsynthese als Ersatz der (eigenen) Stimme wiedergegeben werden sollen [4]; oder sei es, um Textpassagen, die durch Emoticons als "erfreut", "traurig" oder "ironisch" verstanden werden sollen, in angemessener Weise stimmlich einzufärben.

Hier besteht für die Entwickler von benutzer-spezifischer Sprachsynthese die Aufgabe, die genannten Emotionen bzw. Sprechereinstellungen verlässlich voneinander unterscheidbar zu machen, was auch durchaus gelingen kann, wie die Forschung auf diesem Feld zeigt [2][7][8][9].

Literatur

[1] Banse, R. & Scherer, K.R.: Acoustic profiles in vocal emotion expression. In: Journal of Personality and Social Psychology 70 (3), 1996, pp. 614-636.

[2] Burkhardt, F.: Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren. Dissertation Berlin. Aachen, Shaker Verlag: 2001.

[3] Couper-Kuhlen, E.: An Introduction to English Prosody. London: Edward Arnold. 1986.

[4] Iida, A., Campbell, N., Iga, S., Higuchi, F. & Yasumura, M.: A speech synthesis system for assitsing communication. In: Proceedings ISCA Workshop on Speech & Emotion, Nordirland, 2000, pp. 167-172.

[5] Lehiste, I.: The phonetic structure of paragraphs. In: Cohen, A. & Nooteboom, S.G.. (Hg.) Structure and Process in Speech Perception. Berlin: Springer. 1975

[6] Möhler, G. & Mayr, J.: A pitch range model for discourse control. In: ISCA Workshop on Speech Synthesis, Pitlochry, Schottland, 2001.

[7] Murray, I.R. & Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust. Soc. Am. 93 (2), 1993, pp. 1097-1108.

[8] Schröder, M.: Emotional speech synthesis: a review. In: Proceedings Eurospeech 2001.

[9] Schröder, M.: Zur Machbarkeit von Synthese emotionaler Sprache ohne Modellierung der Stimmqualität. Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 1999, Görlitz.

[10] Trouvain, J.: Temposteuerung in der Sprachsynthese durch prosodische Phrasierung. In: Tagungsband 13. Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2002, Dresden, pp. 294-301.

[11] Trouvain, J., Koreman, J., Erriquez, A. & Braun, B.: Artculation Rate Measures and Their Relations to Phone Classification of Spontaneous and Read German Speech. In: Proceedings ISCA Workshop on Adaptation Methods for Speech Recognition, August 2001, Sophia Antipolis, France.