Simulation emotionaler Sprechweise mit konkatenierender Sprachsynthese

Felix Burkhardt und Walter F. Sendlmaier

TU-Berlin, Institut für Kommunikationswissenschaft,
10587 Berlin, Einsteinufer 17 d
email: felixbur@kgw.tu-berlin.de

10. August 1999

Zusammenfassung In dieser Arbeit wird die Möglichkeit der Simulation emotionaler Sprechweise durch konkatenierende Sprachsyntheseverfahren untersucht. Diese bieten in der Regel lediglich Manipulationsmöglichkeiten auf suprasegmenteller Ebene an. In einem Hörversuch wurden synthetisierte Stimuli auf Erkennbarkeit von vier Basisemotionen überprüft. Dabei wurden durch Copy-Synthese generierte Stimuli solchen gegenübergestellt, deren prosodische Merkmalsvariation im Hinblick auf emotionalen Ausdruck durch Regelanwendung bestimmt wurde. Die Auswertung ergab einen signifikanten Unterschied der beiden Stimulusgruppen.

Einleitung

Die Qualität von Sprachsynthesesystemen ist in den letzten Jahren stark verbessert worden. Die Verständlichkeit wie auch die Natürlichkeit kürzerer Textpassagen erreicht bei hochwertigen Systemen annähernd das Niveau von natürlichen sprachlichen Äußerungen.

Als problematisch stellt sich jedoch weiterhin die Natürlichkeit längerer Textpassagen dar. Dies liegt hauptsächlich an der mangelnden Variationsbreite segmenteller wie suprasegmenteller Eigenschaften der synthetisierten Sprachsignale. Die Forderung nach einer größeren Natürlichkeit entspricht somit der Forderung nach stärkerer Variabilität bzgl. des stimmlichen und sprecherischen Ausdrucks. Anhand der Simulation von emotionaler Sprechweise sollen in der vorliegenden Arbeit neue Variationsmöglichkeiten untersucht werden.

Sprachsyntheseverfahren lassen sich auf mehrere Arten klassifizieren, die sich jeweils unterschiedlich zur einer Erhöhung der Variabilität extralinguistischer Merkmale eignen:

- **Systemmodellierende Verfahren** (artikulatorische Synthese) modellieren die Stellungen und Bewegungsabläufe der Artikulatoren und damit die Geometrie des Ansatzrohres durch

mathematische Gleichungen und generieren aus diesen Sprachsignale.

Bedingt durch ungenügende Kenntnis der Zusammenhänge und hohe Komplexität der Modelle ist diese Art der Synthese zwar von hoher Flexibilität, jedoch im Vergleich mit signalmodellierenden Systemen noch von deutlich schlechterer Qualität.

- **Signalmodellierende Verfahren** arbeiten mit der Verkettung von Basiseinheiten. Abgesehen von der Beschaffenheit dieser Einheiten unterscheiden sie sich vor allem durch die Art der verwendeten Kodierung.

  Bei **parametrischer Kodierung** werden die Einheiten in Frames unterteilt, die durch ihre spektralen Eigenschaften beschrieben werden. Dabei wird vorausgesetzt, dass das Signal für die Dauer des Frames als stationär angesehen werden kann. Suprasegmentelle Eigenschaften wie Intonation und Lautdauer sind hierbei explizite Parameter der Dekodierung. Eine Variation segmenteller Eigenschaften lässt sich in der Regel durch Manipulation der Parameter während der Dekodierung erreichen.

  Parametrischen Verfahren liegt das Quelle-Filter-Modell von Fant (1960) zugrunde. Die Synthesequalität ist in der Regel zunächst schlechter als bei Zeitbereichsverfahren. Die Ursachen dafür liegen einerseits an unzulänglicher Parametrisierung des Filters und andererseits an zu einfachen Modellen für das Anregungssignal. Hochqualitative Copy-Synthese ist mit parametrischen Verfahren durchaus möglich.

  Im Gegensatz dazu werden bei einer **Kodierung im Zeitbereich** die Segmente direkt miteinander verkettet, was das Problem der Glättung an den Segmentgrenzen verschärft. Die allgemeine Qualität ist allerdings generell höher als bei parametrischen Verfahren, da ein Großteil an phonetischem Wissen implizit in den Segmenten kodiert ist.

  Die Modellierung der Prosodie ist erst durch neuere Verfahren möglich geworden. Eine Variation segmenteller Eigenschaften zur Synthesezeit ist hierbei grundsätzlich nicht möglich, sondern erfordert eine Erweiterung des Segmentinventars.

Konkatenierende Synthese im Zeitbereich hat in den letzten Jahren gerade im kommerziellen Bereich eine starke Verbreitung gefunden, da sie eine hohe Synthesequalität mit geringem algorithmischen Aufwand verbindet.

Es stellt sich nun die Frage, inwieweit eine Vergrößerung der Variationsbreite stimmlichen Ausdrucks durch Manipulation lediglich solcher Merkmale möglich ist, die sich zur Synthesezeit bei Zeitbereichsverfahren variieren lassen. Weiterhin gibt es grundsätzlich zwei unterschiedliche Ansätze, natürlich wirkende Prosodieparameter für einen zu synthetisierenden Text zu gewinnen:

- **Regelbasierte Algorithmen** wenden auf das Ergebnis einer syntaktischen Analyse des Eingabetextes bestimmte Betonungsregeln an. Eine Erweiterung der Variation des sprecherischen Ausdrucks würde hier die Anwendung gewisser *"Emotionsregeln"* aufgrund einer semantischen Analyse erfordern.

- **Datenbasierte Algorithmen** wenden in der Regel Entscheidungsbaumverfahren auf eine Sprachdatenbank an. Das Ziel ist es dabei, Prosodieparameter zu finden, die in einer syntaktisch möglichst ähnlichen Situation angetroffen wurden. Um hier eine größere Variationsvielfalt zu er-

reichen, müssten die Lernalgorithmen mit unterschiedlichen Datenbanken trainiert werden, da Regeln nicht explizit zur Verfügung stehen.

Daraus ergibt sich eine weitere Fragestellung: Inwieweit unterscheiden sich daten- und regelbasierte Emotionssimulation hinsichtlich ihrer Güte? Die Güte der Simulation soll anhand der Wiedererkennungsrate für eine intendierte Emotion gemessen werden.

Konzeption des Hörversuchs

Aus dem oben skizzierten Problemfeld wird folgenden zwei Fragen explizit nachgegangen.

1. Wie überzeugend lässt sich emotionale Sprechweise mit Zeitbereichs-Syntheseverfahren simulieren?

2. Inwieweit unterscheidet sich dabei daten- von regelbasierter Simulation?

Um sich einer Beantwortung dieser Fragen anzunähern, wurde der Versuch unternommen, stimmlichen emotionalen Ausdruck mit einem Zeitbereichs-Syntheseverfahren zu simulieren. Als Synthesizer wurde das Diphon-Konkatenationssystem MBROLA verwendet (Dutoit et al. (1996)). Zur Gewinnung der Steuerparameter wurden sowohl daten- als auch regelbasierte Verfahren angewendet.

Berücksichtigte Parameter Alle Synthesesysteme bestehen aus einer symbolverarbeitenden (NLP) und einer signalgenerierenden (DSP) Komponente[1]. Die NLP-Komponente erzeugt aus dem Eingabetext eine phonetische Signalbeschreibung und eine Prosodiebeschreibung. Bei Zeitbereichs-DSP Systemen beschränkt sich diese in der Regel auf die Angabe einer Intonationskontur und Lautdauern. Dies erlaubt eine Manipulation suprasegmenteller Merkmale wie Intonationskontur und Betonung. Weiterhin können segmentelle Phänomene wie Silben- oder Lautelision und Assimilation berücksichtigt werden. In eingeschränkter Form ist auch die Umsetzung nichtprosodischer Merkmale möglich; so kann z.B. Jitter durch mikroprosodische Manipulation erzeugt werden.

Nicht möglich sind Variationen auf artikulatorischer oder stimmqualitativer Ebene wie etwa eine Beeinflussung der Phonationsart, der Artikulationsgenauigkeit oder eine Simulation von Nasalierung.

Berücksichtigte Emotionen Um einen Vergleich mit früheren Untersuchungen zu erleichtern, wurden als zu simulierende Emotionen die in der Literatur oft betrachteten vier Basisemotionen *Freude*, *Angst*, *Wut* und *Trauer* verwendet.

Signalgenerierung Als Vorlage für die Stimuli wurden zwei Verfahren angewendet:

- Datenbasierte Simulation
  Um eine datenbasierte Simulation zu realisieren, wurden natürlichsprachliche emotionale Äußerungen hinsichtlich der Parameter Lautdauer, Intonationskontur und Laut-Elision bzw. -Assimilation kopiert. Diese Äußerungen wurden von Schauspielern simuliert und stammen aus einer im Rahmen eines DFG-Projekts erstellten Sprachdatenbank. Dabei wurden von vier Sprechern (zwei männliche

[1] NLP steht für *Natural Language Processing*, DSP für *Digital Speech Processing*.

[2] Die Sätze lauteten: "*An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.*" und "*Das schwarze Blatt Papier befindet sich dort oben neben dem Holzstück.*".

und zwei weibliche) je zwei Sätze[2] als Grundlage für die Erstellung der von MBROLA benötigten Signalbeschreibungsdateien verwendet. Die intendierten Emotionen der Äußerungen waren zuvor in einem Hörtest von mindestens 80 % der Hörer erkannt worden.

- Regelbasierte Simulation
  Um emotionale Sprechweise regelbasiert zu simulieren, wurde der von Burkhardt (1999) beschriebene Prosodiemanipulationsfilter *EmoFilt* verwendet. Dieses Programm ermöglicht die regelhaft beschriebene Manipulation prosodischer Parameter für das MBROLA-Eingabeformat. Beispielsweise lassen sich dadurch Regeln wie: "*Vergrößere den F0-Range um 20 %*" automatisch in eine geänderte Intonationsbeschreibung umsetzen.

  Als Anhaltspunkte für die Manipulationsparameter wurden Ergebnisse aus der Literatur verwendet (z.B. Murray & Arnott (1993); Paeschke et al. (1999)). Die Feineinstellungen wurden von Hand optimiert. Für beide Sätze wurden dieselben Einstellungen verwendet.

  Die Manipulationen für die einzelnen Emotionen sind in Tabelle 1 zusammengefasst. Der Emotionsfilter wurde auf neutrale Versionen der beiden Sätze angewendet, welche von der Phonemisierungskomponente TXT2PHO (Portele (1996)) generiert wurden.

Die Stimuli, bei denen weibliche Sprecher kopiert wurden und die regelbasierten Versionen wurden mit einem weiblichen Diphoninventar synthetisiert, die anderen mit einem männlichen.

Durchführung des Hörtests

Um die Kontextabhängkeit der Hörerurteile zu minimieren, wurden die Stimuli jedem Hörer in einer anderen Zufallsreihenfolge dargeboten.

Es wurde ein Forced-Choice Test gewählt. Nach dem einmaligen Hören des Stimulus musste die Versuchsperson auf die Frage "*Wie klingt der/die Sprecher/in?*" durch die Angabe von *neutral*, *wütend*, *freudig*, *ängstlich* oder *traurig* reagieren.

Insgesamt waren 50 Stimuli zu beurteilen (vier Emotionen plus Neutral mal zwei Sätze mal vier datenbasierte plus der regelbasierten Version). Um die Hörer an das Beurteilungsverfahren und den Syntheseklang zu gewöhnen, wurden vier Stimuli vorangestellt, die bei der Auswertung nicht berücksichtigt wurden.

Teilnehmer waren 10 männliche und 10 weibliche deutsche Muttersprachler im Alter zwischen 21 und 34 Jahren. Die Stimuli wurden in Einzelsitzungen mit Kopfhörern dargeboten.

Auswertung der Ergebnisse

Die Beurteilungen wurden mittels einer mehrfaktoriellen Varianzanalyse mit kompletter Messwiederholung ausgewertet. Die drei Faktoren lauteten: *Emotion* (fünf Ausprägungen), *Sprecher* (fünf Ausprägungen) und *Satz* (zwei Ausprägungen). Es ergaben sich signifikante Effekte für die Faktoren *Emotion* und *Sprecher*. Weiterhin wurde die

Parmeter	Wut	Freude	Angst	Trauer
⌀F0	-	um 40 % angehoben	um 150 % angehoben	um 20 % abgesenkt
F0-Range	um 50 % angehoben	um 40 % angehoben	um 40 % angehoben	um 80 % verringert
F0-Variabilität	um 10 % angehoben	um 100 % angehoben	-	-
F0-Kontur über Phrasen	leicht fallend	-	am Ende steigend	-
starkbetonte Silben	leicht angehoben, Dauer um 10 % verlängert, fallende F0-Kontur	leicht angehoben, Dauer um 20 % verlängert, steigende F0-Kontur	Dauer um 20 % verkürzt	um 20 % verlängert, fallende F0-Kontur
Sprechrate	für unbetonte Silben um 25 % angehoben	für unbetonte Silben um 20 % angehoben	alle Silben um 30 % angehoben	alle Silben um 20 % abgesenkt
Lautdauern	Nasale, Frikative und Plosive um 10 % verkürzt	stimmhafte Frikative um 30 % verlängert	-	-
Sprechpausendauern	-	-	-	um 100 % verlängert
Jitter	3 % Jitter	-	8 % Jitter	10 % Jitter

Tabelle 1: Manipulationen der Stimuli bezogen auf neutrale Sprechweise

Interaktion zwischen *Sprecher* und *Emotion* signifikant. Weder der Faktor *Satz* noch der Innersubjektfaktor *Geschlecht des Beurteilers* wurden signifikant.

In Abbildung 1 sind die durchschnittlichen Bewertungen für die Sprecher in Abhängigkeit von der Emotion sowie die mittlere Bewertung für die einzelnen Emotionen dargestellt. Die Urteile für die beiden Sätze sind dabei zusammengefasst worden.

- **Unterschiede zwischen den Sprechern**
  Signifikante Unterschiede innerhalb der Sprecher ergeben sich zwischen zwei Gruppen: Sprecher 2 und die regelbasierte Version sowie die restlichen Sprecher. Somit sind die regelbasierten Versionen, wenn man von Sprecher 2 absieht, besser erkannt worden als die datenbasierten.

Insgesamt gesehen haben die Schauspieler den emotionalen Ausdruck offensichtlich eher durch artikulatorische oder stimmqualitative Merkmale als durch prosodische Variation erzeugt. Da die regelbasierten Versionen sich ja gerade auf prosodische Manipulationen bezogen, war eine höhere

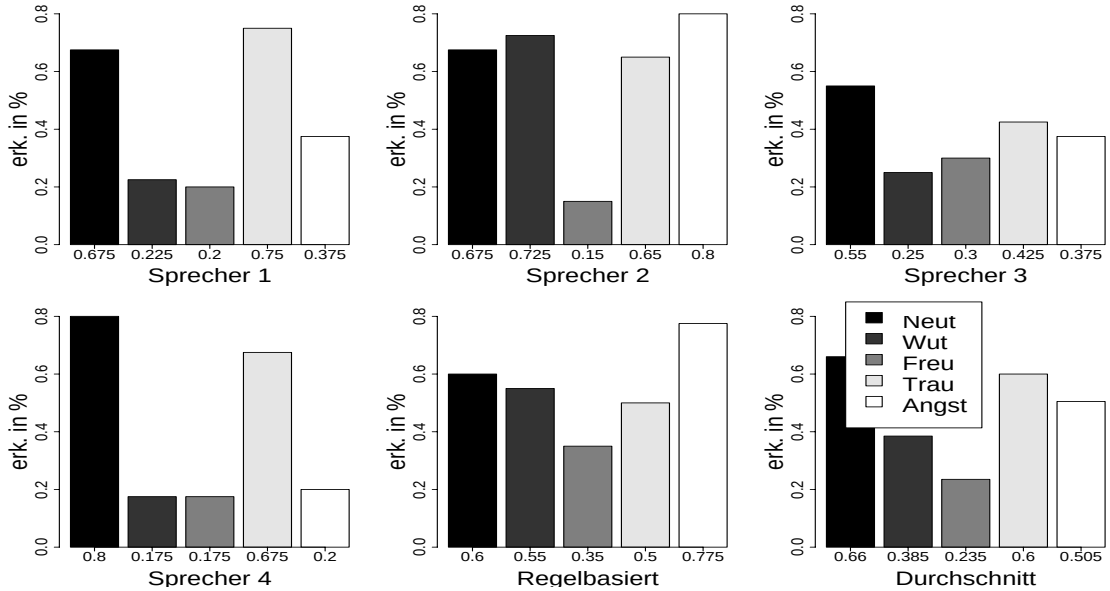

Abbildung 1: durchschnittliche Bewertungen der Stimuli nach Sprechern

Erkennungsrate zu erwarten.

- **Unterschiede zwischen den Emotionen**
  Die Erkennungsrate der Emotionen hat die Rangfolge *neutral*, *traurig*, *ängstlich*, *wütend* und *freudig* . Abgesehen von den Unterschieden zwischen *neutral* / *traurig* und *traurig* / *ängstlich* sind alle Unterschiede signifikant.

  Die freudigen Versionen liegen bei fast allen datenbasierten Versionen unter dem Zufallsniveau von 20 %. Auch bei den regelbasierten Versionen erreicht *Freude* eine Erkennungsrate von lediglich 35 %.

  Die einzigen Emotionen, bei denen die regelbasierten Versionen nicht signifikant besser erkannt wurden als die datenbasierten, sind *Trauer* und *Neutral*.

  Bei den *Angst* simulierenden regelbasierten Stimuli ist die Erkennungsrate auffällig hoch. Anscheinend waren die zugrundeliegenden Regeln besonders adäquat.

- **Konfusionsmatrizen**

In Abbildung 2 sind Konfusionsmatrizen für die daten- und die regelbasierten Simulationen dargestellt.

Bei der Konfusionsmatrix für die datenbasierten Stimuli fällt auf, dass allgemein die Emotionen häufig als *Neutral* klingend beurteilt wurden. Vermutlich kam hier der Effekt zum Tragen, dass bei Unsicherheit des Beurteilers für *Neutral* entschieden wurde. Ansonsten gibt es die größten Verwechslungen für *Freude*. Sie wurde am häufigsten mit *Angst* und am seltensten mit *Wut* verwechselt.

Bei den regelbasierten Versionen wurde *Wut* am häufigsten mit *Neutral* verwechselt und *Freude* mit *Wut*. Diese Verwechslung läßt sich durch eine ähnliche Erregungstufe der beiden Emotionen erklären. Alle anderen Verwechslungen lagen unter dem Zufallsniveau.

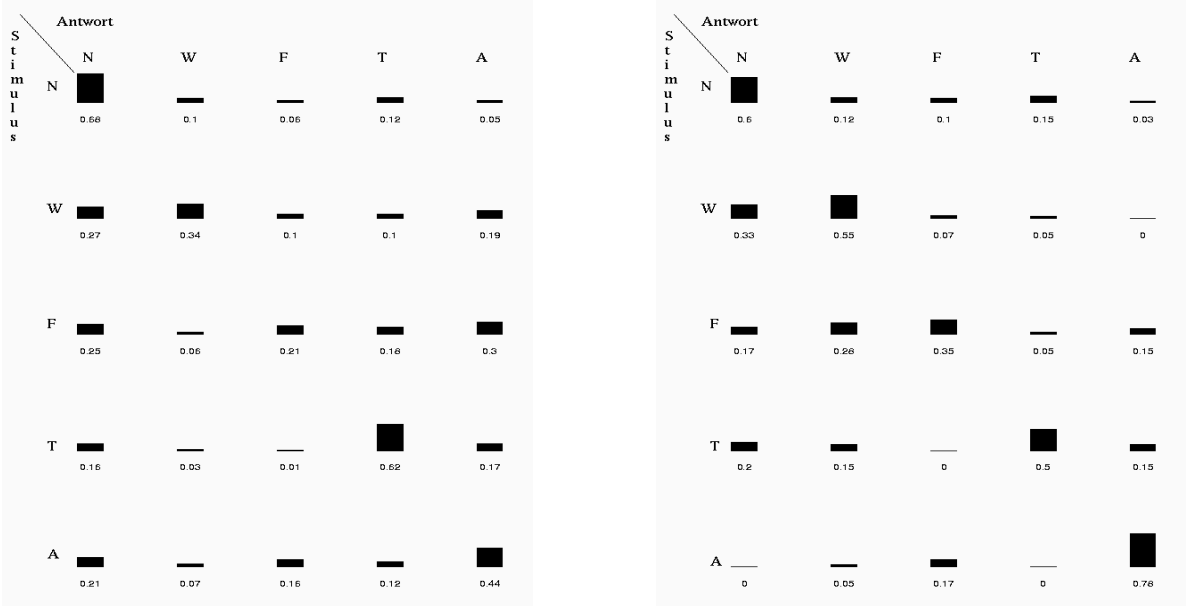

Abbildung 2: Konfusionsmatrix für die datenbasierten (links) und die regelbasierten (rechts) Stimuli. Die Zeilen entsprechen den intendierten Emotionen, die Spalten den erkannten.

Bezug der Ergebnisse zu früheren Untersuchungen Ähnliche Untersuchungen liegen von Cahn (1989); Carlson et al. (1992); Vroomen et al. (1993); Murray & Arnott (1995); Heuft et al. (1998); Rank & Pirker (1998); Schröder (1999) vor.

Gemeinsam ist fast allen Untersuchungen, dass verschiedene Emotionen sehr unterschiedlich gut erkannt wurden. Weiterhin liegen bei fast allen Untersuchungen die Erkennungsraten ähnlich hoch (im Schnitt zwischen 30 und 50 %). Die einzige Ausnahme ist hier Vroomen et al. (1993), bei der von überdurchschnittlich hohen Erkennungsraten berichtet wird. Eventuell eignet sich das Holländische besonders gut zur Simulation emotionaler Sprechweise durch die Variation prosodischer Merkmale.

Auch in anderen Untersuchungen wurde Freude eher schlechter (Carlson et al. (1992); Murray & Arnott (1995); Heuft et al. (1998); Schröder (1999)) und Trauer eher besser (Cahn (1989); Murray & Arnott (1995); Rank & Pirker (1998); Schröder (1999)) erkannt. Scherer (1981) weist darauf hin, das negative Emotionen anscheinend generell besser erkannt werden als positive.

Es findet sich allerdings relativ wenig Übereinstimmung zwischen den Ergebnissen dieser Untersuchung und früheren Ergebnissen in Bezug auf Verwechslungen zwischen den Emotionen. Offensichtlich sind die sprecherspezifischen Unterschiede im Bereich emotionaler stimmlicher Ausdrucksstrategien sehr groß.

Ein Vergleich zwischen verschiedenen Untersuchungen wird selbstverständlich dadurch erschwert, dass verschiedene Syntheseverfahren angewendet wurden, unterschiedlich viele Basisemotionen untersucht wurden und die Stimuli sowohl vom Inhalt als auch von der Sprache her unterschiedlich waren.

Diskussion und Ausblick

Die Ergebnisse zeigen, dass emotionale Sprechweise durch die beschränkten Möglichkeiten einer Zeitbereichssynthese erkennbar zu simulieren ist. Dabei gibt es allerdings beträchtliche Unterschiede zwischen einzelnen Emotionen.

Regelbasierte Modellierung ergibt dabei generell bessere Ergebnisse, da die Regeln speziell auf die Möglichkeiten des Synthesizers abgestimmt werden können.

Die großen Unterschiede zwischen den Sprechern deuten auf unterschiedliche Sprecherstrategien hin, Emotionen auszudrücken (vgl. auch Schröder (1999)).

Die erhebliche Diskrepanz der Erkennungsrate zwischen resynthetisierten und natürlichen Stimuli legt nahe, dass zur Simulation nicht nur erkennbarer, sondern auch natürlich wirkender emotionaler Sprechweise eine Modellierung stimmqualitativer und artikulatorischer Merkmale unumgänglich ist.

Insofern scheint die Verwendung von Synthesizern, die eine Variation segmenteller Merkmale ermöglichen, attraktiv. Bei Carlson et al. (1992) z.B. stieg die Erkennungsrate deutlich für Stimuli, bei denen auch das Anregungssignal eines Formantsynthesizers dem natürlichen Vorbild nachempfunden war.

Zukünftige Forschung im Rahmen dieses Projektes wird sich mit der Simulation emotionaler Sprechweise durch Formantsynthese beschäftigen.

Danksagung

Diese Untersuchung wurde von der Deutschen Forschungsgemeinschaft gefördert, Kennziffer Se462/3-1.

Literatur

F. Burkhardt. Simulation des emotionalen Sprecherzustands ‚Freude' mit einem Sprachsyntheseverfahren. Magisterarbeit, TU-Berlin, Institut für Kommunikationswissenschaft, 1999.

J. E. Cahn. The Affect Editor. *Journal of the American Voice I/O Society*, 8:1–19, 1989.

R. Carlson, G. Granström & L. Nord. Experiments With Emotive Speech, Acted Utterances And Synthesized Replicas. *Speech Communication*, 2:347–355, 1992.

T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. Van der Vreken. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proc. ICSLP'96, Philadelphia*, 3:1393–1396, 1996.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, 's-Gravenhage, 1960.

B. Heuft, T. Portele & M. Rath. Emotions in Time Domain Synthesis. *Proc. ICSLP*, 1998.

I. R. Murray & J. L. Arnott. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *JASA*, 2:1097–1107, 1993.

I. R. Murray & J. L. Arnott. Implementation and Testing of a System for Producing Emotion-by-rule in Synthetic Speech. *Speech Communication*, 16:369–390, 1995.

A. Paeschke, M. Kienast & W. F. Sendlmeier. F0-Konturs in Emotional Speech. In *Proc. ICPhS, San Francisco '99*, 1999.

T. Portele. Das Sprachsynthesesystem Hadifix. *Sprache und Datenverabeitung*, 21:5–22, 1996.

E. Rank & H. Pirker. Generating Emotional Speech with a Concatenative Synthesizer. In *Proc. of the ICSLP98*, 1998.

K. R. Scherer. Speech and Emotional States. In J. K. Darby, Hrsg., *Speech Evaluation in Psychiatry*, Seite 189–218. Grune & Stratton, 1981.

M. Schröder. Can Emotions be synthesized without Controlling Voice Quality? In *Phonus 4, Forschungsbericht Institut für Phonetik, Universität des Saarlandes*, 1999.

J. Vroomen, R. Collier & S. Mozziconacci. Duration and Intonation in Emotional Speech. In *Eurospeech 93, Berlin*, Band 1, Seite 577–580, 1993.