Textgeführte Generierung von Sprechbewegungen

Bernd J. KRÖGER
Klinik für Phoniatrie, Pädaudiologie
und Kommunikationsstörungen
Universitätsklinikum der RWTH Aachen
Pauwelsstr. 30, 52074 Aachen, Germany
bkroeger@ukaachen.de

Christiane NEUSCHAEFER-RUBE
Klinik für Phoniatrie, Pädaudiologie
und Kommunikationsstörungen
Universitätsklinikum der RWTH Aachen
Pauwelsstr. 30, 52074 Aachen, Germany
cneuschaefer@ukaachen.de

Zusammenfassung

Es wird ein Modell zur Generierung und Visualisierung von Sprechbewegungen vorgestellt, das als Hilfsmittel in der Therapie von Sprechstörungen zum Einsatz kommt. Um dem Therapeuten die intensive Beschäftigung mit phonetischer Lautschrift zu ersparen, können mittels dieses Ansatzes artikulatorische Bewegungsabläufe von Wörtern und Sätzen textgeführt, d.h. aufgrund der Eingabe einer Graphemfolge generiert und visualisiert werden.

Einführung

Es existieren heute bereits viele Systeme zur textgeführten Generierung von akustischen Sprachsignalen (text-to-speech, TTS, Paulus 1998, siehe auch Angabe von Internetadressen bei Campbell et al. 2001), wobei einige dieser Systeme bereits einen sehr hohen Qualitätsstandard erreicht haben. Bei artikulationsbasierten Systemen (Coker et al. 1973, Mermelstein 1973, Rubin et al. 1981, Saltzman & Munhall 1989, Gabioud 1994, Wilhelms-Tricarico 1995, Hoole 1999, Beautemps et al. 2001) stellte sich die Frage nach einer graphemischen Eingabe bislang nicht, da die Sprachqualität dieser Systeme noch nicht an die Qualität der in textgeführten Systemen Verwendung findenden, akustisch basierten Synthesetechniken heranreicht (Kröger 1998).

1 Die Funktionen des Modells

Zentraler Baustein des Modells ist die Generierung von Sprechbewegungen (Artikulationsbewegungen; Bewegungen von Lippen, Zunge, Gaumensegel etc. beim Sprechen) und die nachfolgende Visualisierung der Bewegungen im Mediosagittalschnitt als Filmsequenz. Die Benutzerschnittstellen wurden möglichst komfortabel gestaltet. Die Eingabe der Wörter bzw. der kurzen Sätze kann deshalb graphemisch, per Lautschrift oder durch das Auswählen des Wortes bzw. des Satzes aus einer zuvor markierten Liste erfolgen. (Wort-)Listen existieren bereits zu vielen Alltagsthemen, insbesondere zu Alltagsthemen rund um Krankenhaus und Rehabilitation. Jeder Eintrag einer Liste liegt bereits graphemisch und in Lautschrift vor. Die zugehörigen Sprechbewegungen wurden darüber hinaus mit natürlichsprachlichen akustischen Signalen synchronisiert (Kröger & Neuschaefer-Rube 2002).

2 Die Ebenen des Modells

Bei der textgeführten Generierung von Sprechbewegungen können drei Ebenen differenziert werden (Tab. 1). Die Ebene der Lautschrift wird in unserem System als allophonische Transkription definiert. Einerseits erlauben Allophone - in Analogie zu Phonemen - die Einführung des Prinzips der phonetischen Unterspezifikation (in Analogie zur phonologischen Unterspezifikation in Merkmal-basierten Phonologien, siehe Keating 1988, Farnetani & Recasens 1999). Unterspezifikation auf der phonetischen Ebene - im folgenden auch als „artikulatorische Unterspezifikation“ bezeichnet - bedeutet, dass auf dieser Ebene noch nicht alle artikulatorischen Details eines Lautes spezifiziert sein müssen und führt in direkter Weise zur Generierung von Koartikulation (siehe Kapitel 5). Andererseits sind Allo-

phone aber dergestalt phonetisch konkret, dass auf dieser Ebene bereits einige sprecherspezifische Faktoren (z.B. sprecherspezifische /r/-Realisierung) und beispielsweise auch Sprechtempo- und Sprechstil-abhängige Faktoren (z.B. Grad der Verschleifungen) definiert sein müssen.

Die Ebene der Sprechbewegungen ist zweistufig: Zunächst werden artikulatorische Bewegungsverläufe für definierte artikulatorische Parameter erstellt (Abb. 1; zur Diskussion von Systemen artikulatorischer Modell-Parameter siehe Harshman et al. 1977, Browman & Goldstein 1990, Hoole 1999, Kröger 2000, Beautemps 2001). Nachfolgend werden im 40ms-Takt mediosagittale Schnittbilder generiert (Abb. 2) und als Filmsequenz visualisiert (siehe auch Kröger et al. 2000, Kröger 2001).

Ebene	Erläuterung	Beispiel
Rechtschrift	Graphemfolge eines Textes	„Campus“
Lautschrift	Allophonische Transkription	[kampUs]
Sprechbewegungen	Bewegungsverläufe der Artikulatoren	Abb. 1
	Mediosagittalbilder (Filmsequenz)	Abb. 2

Tabelle 1. Ebenen zur textgeführten Generierung von Sprechbewegungen

3 Generierung der Lautschrift

Die in unserem Modell realisierte Generierung von Lautschrift basiert auf einem System von Transkriptionsregeln in Kombination mit einem Affixverzeichnis (Verzeichnis gebundener Morpheme: Präfixe, Suffixe und Infixe), einem Verzeichnis von Funktionswörtern (Artikel, Pronomina, Präpositionen, Konjunktionen, Adverbien, Funktionsverben/Formverben) und einem Ausnahmelexikon für Wortstämme und deren Alternanten, die durch die Regeln und durch das Affix- und Funktionswörterverzeichnis nicht erfasst werden. Das Ausnahmelexikon kann vom Benutzer eingesehen und erweitert werden. Zur Zeit werden noch in ca. 10% aller Fälle fehlerhafte Transkriptionen erzeugt. Die Generierung einer Transkription ist aber auch in diesen Fällen sinnvoll, da der Therapeut nun bereits über eine Grundtranskription des eingegebenen Wortes verfügt, die nur noch korrigiert werden muss.

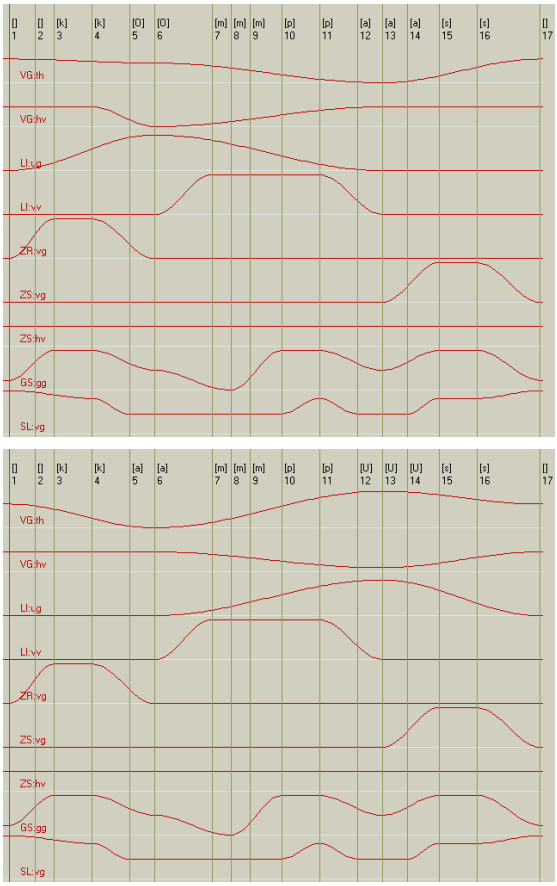

Abbildung 1. Artikulatorische Bewegungsverläufe der Wörter „Kompass“ (oben) und „Campus“ (unten). *Senkrecht:* Artikulatorische Parameter von oben nach unten: VG:th (vokalische Gesamtformung tief-hoch), VG:hv (vokalische Gesamtformung hinten-vorne), LI:ug (Lippen ungerundet-gerundet), LI:vv (Lippen vokalisch-verschlossen), ZR: vg (Zungenrücken vokalisch-gehoben), ZS:vg (Zungenspitze vokalisch-gehoben), ZS:hv (Zungenspitze hinten-vorne), GS:gg (Gaumensegel gesenkt-gehoben), SL:vg (Stimmritze verschlossen-geöffnet). *Waagerecht:* Die Kurven geben die Verläufe der artikulatorischen Parameter an. Die Zeit verläuft dabei von links nach rechts. Die senkrechten Linien kennzeichnen artikulatorisch definierte Anfangs-Mittel- und End-Zeitpunkte der Laute (Lautlabel). Label 1 und 2: Präphonation; Label 3 und 4: Beginn bzw. Ende der dorsalen Verschluss-

bildung des [k]; Label 5: Beginn der vokalischen Phonation des ersten Vokals; Label 6: Vokalmittelpunkt des ersten Vokals; Label 7 und 9: Beginn bzw. Ende des labialen Verschlussbildung des [m]; Label 8: maximale Absenkung des Gaumensegels; Label 10 und 11: Beginn bzw. Ende der labialen Verschlussbildung des [p]; Label 12 und 14: Beginn bzw. Ende der vokalischen Phonation des zweiten Vokals; Label 13: Vokalmittelpunkt des zweiten Vokals; Label 15 und 16: Beginn und Ende der apikalen Engebildung des [s]; Label 17: Postphonation.

Abbildung 2. Mediosagittales Schnittbild des Vokals [a]. Der Punkt oberhalb der Stimmlippen kennzeichnet den Entstehungsort des Phonationsschalls.

4 Generierung der Sprechbewegungen

Bei der Generierung der Sprechbewegungen ist die Idee der artikulatorischen Unterspezifikation von zentraler Bedeutung. Jedes Allophon ist phonetisch nur minimal spezifiziert. So werden beispielsweise labiale Konsonanten (z.B. [m] und [p]) abgesehen von der Einstellung des Gaumensegels und der Stimmlippen nur bezüglich der Einstellung eines labialen Verschlusses festgelegt. Insbesondere wird die Lage des Zungenrücken (und der Zungenspitze) für diese Konsonanten nicht angegeben. Diese Laute sind somit bezüglich der Einstellung bzw. bezüglich der Bewegung von Zungenrücken und Zungenspitze „artikulatorisch und koartikulatorisch frei“. Die Einstellung von Zungenrücken und Zungenspitze wird erst innerhalb der artikulatorischen Realisierung der gesamten Lautfolge aufgrund von Vorgaben durch die Nachbarlaute - insbesondere aufgrund benachbarter Vokale - spezifiziert.

Anhand der Beispielwörter „Kompass“ und „Campus“ (Abb. 1) wird deutlich, wie ein unterschiedlicher vokalischer Kontext zu unterschiedlicher Realisierung von Konsonanten führt. So tritt während der konsonantischen labialen Verschlussphase von [m] und [p] in „Kompass“ eine Senkung und Vorverlagerung des Zungenrücken, in „Campus“ jedoch eine Hebung und Rückverlagerung des Zungenrücken auf (siehe die Verläufe der Parameter der vokalischen Gesamtformug in Abb. 1. Der Zeitbereich des bilabialen Verschlusses des [m] und [p] liegt zwischen Lautlabel 7 und 11.)

Diskussion und Ausblick

Das hier vorgestellte Modell zur textgeführten Generierung von Sprechbewegungen ermöglicht die Generierung von Sprechbewegungen ohne explizite Eingabe von phonetischer Lautschrift. Dies ist beim Einsatz des Systems im Bereich der Therapie von Sprechstörungen aus rein praktischer Sicht sinnvoll, da es dem Therapeuten die intensive Auseinandersetzung mit phonetischer Lautschrift erspart.

Zur Generierung der Allophonfolge wird angestrebt, neben dem hier verwendeten Kombination aus Regelsystem, Affix-, Funktionswörterverzeichnis und Ausnahmelexikon alternative Ansätze (z.B. statistische Modelle (Wolff & Eichner 2001) oder neuronale Modelle (Mana et al. 2001) anzubieten.

Aufgrund erster Erfahrungen im klinischen Alltag kann festgestellt werden, dass das hier vorgestellte Modell das Artikulationstraining für diejenigen sprechgestörten Patienten erleichtert, die in der Lage sind, die mediosagittale Darstellung der Sprechorgane kognitiv zu erfassen. Insbesondere ist die Visualisierung der Bewegungen der normalerweise nicht sichtbaren Arti-

kulationsorgane (Gaumensegel, hinterer Zungenbereich, Stimmlippen) für die Patienten sehr hilfreich.

Literatur

Beautemps D., Badin P. & Bailly G. (2001) *Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modelling.* Journal of the Acoustical Society of America 109, pp. 2165-2180.

Browman C.P. & Goldstein L. (1990) *Tiers in articulatory phonology, with some implications for casual speech.* In: "Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech." J. Kingston & M. E. Beckman, eds., Cambridge University Press, Cambridge, pp. 341-376. Auch in: Haskins Laboratories Status Report on Speech Research SR-92 (1987), pp. 1-30.

Campbell N., Hess W., Möbius B. & van Santen J. (2001) *The ISCA special interest group in speech synthesis.* Proceedings of the European Conference on Speech Communication and Technology, Proc. EUROSPEECH 2001, pp. 1149-1152.

Coker C. H., Umeda N. & Browman C.P. (1973) *Automatic synthesis from ordinary english text.* IEEE Transactions on Audio and Electroacoustics AU-21, pp. 293-298.

Farnetani E. & Recasens D. (1999) *Coarticulation models in recent speech production theories.* In "Coarticulation: theory, data and techniques", W. J. Hardcastle & N. Hewlett, eds., Cambridge University Press, Cambridge, pp. 31-68.

Gabioud B. (1994) *Articulatory models in speech synthesis.* In "Fundamentals of Speech Synthesis and Speech Recognition. Basic Concepts, State of the Art, and Future Challenges", E. Keller & J. Caelen, eds., John Wiley, Chichester, pp. 215-230

Harshman R., Ladefoged P. & Goldstein L. (1977) *Factor analysis of tongue shapes*. Journal of the Acoustical Society of America 62, pp. 693-707.

Hoole P. (1999) *On the lingual organization of the German vowel system.* Journal of the Acoustical Society of America 106, pp. 1020-1032.

Keating P. A. (1988) *Underspecification in phonetics.* Phonology 5, pp. 275-292.

Kröger B. J. (1998) *Ein phonetisches Modell der Sprachproduktion.* Niemeyer Verlag, Tübingen.

Kröger B. J. (2000) *Analyse von MRT-Daten zur Entwicklung eines vokalischen Artikulationsmodells auf der Ebene der Areafunktion.* In „Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation 20", K. Fellbaum, ed., w.e.b.-Univ.-Verl., Dresden, pp. 201-208.

Kröger B. J., Winkler R., Mooshammer C. & Pompino-Marschall B. (2000) *Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results.* Proceedings of 5th Seminar on Speech Production: Models and Data (Kloster Seeon, Bavaria), pp. 333-336.

Kröger B. J. (2001) *Das funktionale Artikulationsmodell FARM: Modellierung von zeitlicher und räumlicher Koartikulation*. In „Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation 22", W. Hess & K. Stöber, eds., w.e.b.-Univ.-Verl., Dresden, pp. 123-130.

Kröger B. J. & Neuschaefer-Rube C. (2002) *Eine Datenbank zur artikulatorisch-akustischen Synchronisation.* Tagungsband DAGA 2002, Bochum.

Mana F., Massimino P. & Racchiotti A. (2001) *Using machine learning techniques for grapheme to phoneme transcription.* Proceedings of the European Conference on Speech Communication and Technology, Proc. EUROSPEECH 2001, pp. 1915-1918.

Mermelstein P. (1973) *Articulatory model for the study of speech production.* Journal of the Acoustical Society of America 53, pp. 1070-1082.

Paulus E. (1998) *Sprachsignalverarbeitung.* Spektrum Akademischer Verlag, Heidelberg Berlin.

Rubin P., Bear T. & Mermelstein P. (1981) *An articulatory synthesizer for perceptual research*. Journal of the Acoustical Society of America 77, pp. 640-648.

Saltzman E.L. & Munhall K.G. (1989) *A dynamic approach to gestural patterning in speech production*. Ecological Psychology 1, pp. 333-382.

Wilhelms-Tricarico R. (1995) *Physiological modelling of speech production: Methods for modelling soft-tissue articulators*. Journal of the Acoustical Society of America 97, pp. 3085-3098.

Wolff M. & Eichner, M. (2001) *Untersuchungen zum statistischen Zusammenhang zwischen orthographischer und phonetischer Repräsentation deutscher Wörter.* In „Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation 22", W. Hess & K. Stöber, eds., w.e.b.-Univ.-Verl., Dresden, pp. 315-322.