EMA-Daten und akustische Daten als Basis für artikulatorische Resynthese

Bernd J. Kröger[1] und Natalie Koster[2]
[1]Institut für Deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin
[2]Institut für Phonetik der Universität zu Köln

1 Einleitung

Resynthese (Kopie-Synthese natürlich gesprochener ganzer Äußerungen) ist hilfreich zur Abschätzung der maximal erreichbaren Qualität eines Synthetisators. Im Falle artikulatorischer Sprachsynthese können neben akustischen Daten auch artikulatorische Daten als Basis für eine Resynthese benutzt werden. Die artikulatorischen Daten wurden im Rahmen dieser Untersuchung mittels der Methode der elektromagnetischen Artikulographie (EMA) gewonnen.

Während zunächst ein auf akustischen Daten und auf unserem segmentalen Steuermodell basierendes Resyntheseverfahren entwikkelt wurde (Kröger et al. 1997) steht bei dem im folgenden vorgestellten Verfahren die Einbeziehung insbesondere artikulatorischer Daten im Mittelpunkt. Es wurde eine kleine Datenbasis (acht Äußerungen, zwei Sprecher) resynthetisiert.

2 Die artikulatorischen Daten

Die EMA-Daten liefern die Auslenkung der Unterlippe, der Zungenspitze und von zwei Punkten im vorderen Bereich des Zungenrükken als Funktion der Zeit in der jeweiligen Hauptbewegungsrichtung des Artikulators für die entsprechende Äußerung (Abb. 1). Zunächst werden kinematische Daten extrahiert: Zur Festlegung der Dauer einzelner artikulatorischer Gesten werden die Zeitpunkte aller Minima und Maxima der vier Auslenkungskurven bestimmt. Anhand der Daten des Zungenrücken (ZR1 und ZR2) können damit Anfangs- und Endzeitpunkte der vokalischen und der dorsalen konsonantischen Gesten, anhand der Daten der Zungenspitze (ZS) die der apikalen konsonantischen Gesten und anhand der Daten der Unterlippe (LI) die der labialen konsonantischen Gesten abgeschätzt werden (Eine explizite Aufstellung der Gesten des Beispiels von Abb. 1 ist in Tab. 1 gegeben). Neben diesen kinematischen Parametern können mittels des von uns entwickelten gestischen Analyseverfahrens (Kröger et al. 1995) für jede Geste auch dynamische Parameter (virtuelles Target, Eigenperiodendauer, Phasenwert des Gestenendes) ermittelt werden. Das virtuelle Target repräsentiert den von der Geste angestrebten räumlichen Zielpunkt. Der Parameter Eigenperiodendauer definiert das Zeitintervall, in dem eine Geste das (virtuelle) Target annähernd erreicht und ist ein Maß für die Artikulatorgeschwindigkeit. Phasenwerte beschreiben die Artikulator-Target-Distanz zum durch den Phasenwert definierten Zeitpunkt. Sie sind ein Maß für den zu diesem Zeitpunkt vorliegenden Grad der Realisierung einer Geste. Der Endphasenwert gibt somit den von einer Geste insgesamt erreichten Realisierungsgrad an (Kröger 1993). Die so erhaltenen artikulatorischen Daten können als Basis für Resynthesen mittels unseres gestischen Produktionsmodells (Kröger 1993) genutzt werden.

Allerdings sind die ausschließlich anhand von EMA-Daten durchgeführten Resynthesen wenig zufriedenstellend. Hierfür können zwei Gründe angeführt werden.

Erstens: Die akustisch bzw. artikulatorisch gemessenen Lautintervalle stimmen nicht überein. Ein Vergleich der akustischen und der artikulatorischen Daten zeigt, daß anhand der artikulatorischen Daten nicht immer sicher der Zeitpunkt des Beginns und des Endes einer konsonantischen Verschlußbildung abgelesen werden kann. Beispielsweise führt die Abschätzung des Zeitpunktes der Verschlußlösung des „d“ in „das“ anhand der artikulatorischen Daten (Abb. 1, Marke V2) und anhand der akustischen Daten (Abb. 1, Marke V1) zu unterschiedlichen Ergebnissen. Grund hierfür ist, daß die EMA-Technik nur die Bewegung *eines* Oberflächenpunktes des jeweiligen Artikulators, nicht aber die Bewegung, Kontaktbildung und Formänderung des gesamten Artikulators erfassen kann. Liegt der gemessene Oberflächenpunkt des Artikulators nicht genau im Zentrum der Verschlußbildung, so erscheint die gemessene Verschlußbildung immer als zu kurz, da die zur Konstanz der artikulatorischen Auslenkungskurve führende Kontaktbildung mit einer harten Sprechtraktbegrenzung (Gaumen oder Zahndamm) zeitlich nach der Verschlußbildung beginnt und vor der Verschlußlösung endet.

Zweitens: Anhand der vorgegebenen EMA-Daten liegt keine Information über die Artikulatorbewegungen des Velums und der Glottis vor. Damit fehlt insbesondere die Information über den Öffnungsgrad zwischen Nasen- und Rachenraum (zur Bestimmung des Grades der Nasalität), über den Abstand der glottalen Stellknorpel zueinander (zur Bestimmung des Grades der Stimmhaftigkeit) und über die Stimmlippenspannung (zur Bestimmung der Grundfrequenz).

3 Das Verfahren der gestischen Resynthese

In dem von uns entwickelten Verfahren der gestischen Resynthese werden deshalb neben den eben beschrieben artikulatorisch-kinematischen Daten (Anfangs- und Endzeitpunkt der Gesten, siehe Abb. 1) und neben den artikulatorisch-dynamischen Daten (virtuelles Target, Eigenperiodendauer und Endphasenwert) auch akustische Daten verwendet, nämlich die anhand von Oszillogramm und Sonagramm ermittelbaren Zeitpunkte der oralen Verschlußbildung und –lösung bei Konsonanten und die anhand von Oszillogramm und Sonagramm ermittelbaren Zeitpunkte des Beginns von stimmhaften bzw. stimmlosen Signalabschnitten.

Die Resyntheseprozedur besteht aus sechs Schritten: (1) Die Eigenperiodendauer aller vokalischen und konsonantischen Gesten wird anhand der artikulatorischen Meßdaten gesetzt. (2) Die Endphasenwerte der vokalischen Gesten werden so angesetzt, daß die Gestendauer innerhalb der Resynthese genau der gemessenen artikulatorisch-kinematischen Gestendauer entspricht. Es wird somit davon ausgegangen, daß die vokalischen Gesten ohne zeitliche Lücke, aber auch ohne zeitliche Überlappung aneinandergereiht werden, so daß keine weitere Berechnung zur zeitlichen Koordinierung der vokalischen Gesten untereinander durchgeführt werden muß. Die zeitliche Lage aller vokalischen Gesten ist damit definiert. (3) Die Endphasenwerte der konsonantischen Gesten werden so angesetzt, daß das Gestenende innerhalb der Resynthese genau mit dem akustisch gemessenen Zeitpunkt der konsonantischen Verschlußlösung des von der Geste realisierten Lautsegments übereinstimmt. (4) Die zeitliche Lage der konsonantischen Gesten wird relativ zu den Anfangszeitpunkten der vokalischen Gesten in Form der Assoziationsphasenwerte quantifiziert. Die Assoziationsphasenwerte können bei bekanntem Anfangszeitpunkt der konsonantischen Gesten (nach den artikulatorischen Daten) und bekannter Eigenperiodendauer direkt berechnet werden. (5) Die Velumsöffnung (Glottisöffnung) wird bei Nasalen (Frikativen und Plosiven) so justiert, daß dieses Öffnungsintervall das Zeitintervall der oralen Verschlußbildung des Nasals (Frikativs oder Plosivs) überdeckt. (6) Der Grundfrequenzverlauf und der Intensitätslevelverlauf des akustischen Signals werden durch Variation von Stimmlippenspannung und glottalem Öffnungsgrad justiert (siehe Kröger et al. 1997). Damit sind alle gestischen Parameter determiniert, und die Resynthese kann durchgeführt werden.

4 Ergebnisse

Ein Vergleich der mittels unseres Analyseverfahrens (Kröger et al. 1995) aus den artikulatorischen Daten gemessenen Endphasenwerte mit den hier anhand der kinematischen und akustischen Daten für die Resynthesen angesetzten Endphasenwerte zeigt, daß die artikulatorisch gemessenen Werte immer kleiner als die für die Resynthese angesetzten Werte sind. Dies ist eine Folge der systematisch falsch artikulatorisch gemessenen absoluten Zeitpunkte für die Verschlußlösung (siehe oben).

Der Vergleich der aus den artikulatorischen Daten berechneten Assoziationsphasenwerte mit dem vom gestischen Produktionsmodell (Kröger 1993) bisher angenommen Wert zeigt, daß die anhand der Daten berechneten Werte geringer sind als der von der zeitlichen Koordinierungsregel des Produktionsmodells angenommene Wert. Während in unserem Produktionsmodell bisher angenommen wurde, daß der Anfangszeitpunkt der von einer konsonantischen Geste erzeugten Konstriktion immer mit dem Anfangszeitpunkt der nachfolgenden vokalischen Geste übereinstimmt, zeigen die artikulatorischen Daten, daß der Anfangszeitpunkt einer konsonantischen Konstriktion (z. B. Anfangszeitpunkt des Verschlusses bei Plosiven und Nasalen, des Beinahverschlusses bei Frikativen) im Vergleich zum Anfangszeitpunkt der nachfolgenden vokalischen Geste *später* liegen kann (siehe auch Löfqvist und Gracco 1999). Die zeitliche Verzögerung kann dabei bis zu 20% der Gesamtdauer der nachfolgenden Vokalgeste ausmachen. Der Grund für diese Verzögerung liegt darin, daß der Zungenrücken im zeitlichen Anfangsbereich einer vokalischen Geste noch vollständig im Targetbereich der vorhergehenden vokalischen Geste liegt und somit auch dieser zeitliche Anfangsbereich der nachfolgenden Vokalgeste noch zur Wahrnehmung der Vokalqualität des vorhergehenden Vokals beiträgt.

Allerdings beeinflußt die Koartikulation zwischen konsonantischen und vokalischen Gesten die Lage der kinematischen Maxima und Minima beider Gesten, so daß von der kinematischen zeitlichen Ausdehnung einer Geste nicht in direkter Weise auf die zugrundeliegende (dynamische) Ausdehnung der vokalischen Gesten geschlossen werden kann. Beispielsweise liegt der gemessene Anfangszeitpunkt der Vokalgeste des „a“ in „das“ (Abb. 1, Marke 2) aufgrund des koartikulatorischen Einflusses der apikalen Verschlußgeste des „d“ sehr spät.

5 Diskussion

Es konnte ein Verfahren zur gestischen Resynthese ganzer Äußerungen anhand von artikulatorischen und akustischen Daten entwickelt werden. Die mittels der EMA-Technik vorliegenden Daten sind allerdings für die Erstellung gestischer Resynthesen nicht hinreichend. Dies liegt vor allem daran, daß diese Technik nur die Bewegung einzelner Oberflächenpunkte, nicht aber die Formung und Bewegung des gesamten Artikulators liefert.

Allein anhand von akustischen Daten können *gestische* Resynthesen aber auch nicht erstellt werden. Insbesondere liefern die EMA-Daten Aufschluß über die zeitliche Lokalisierung und das kinematische Profil der Transitionsphasen aller Gesten. Nur die artikulatorischen Daten geben direkten Aufschluß über die zumeist hinter konsonantischen Verschlußbildungen versteckten vokalischen Artikulatorbewegungen und über die innerhalb der akustischen Vokalsegmente auftretenden konsonantischen Transitionsbewegungen. Insbesondere kann die genaue zeitliche Überlappung von vokalischen und konsonantischen Gesten (Grad der gestischen Koproduktion) auch nur anhand der artikulatorischen Daten determiniert werden.

6 Literatur

Kröger, B.J. (1993): „A gestural approach for controlling an articulatory speech synthesizer“, *Proceedings of the European Conference on Speech Communication and Technology, EUROSPEECH '93*, **Vol. 3**, 1903-1907.

Kröger, B.J., Opgen-Rhein, C., Sachse, G. (1997): „Artikulatorische Resynthese anhand akustischer Daten mittels eines segmentalen Produktionsmodells“, in: *Fortschritte der Akustik: Plenarvorträge und Fachbeiträge der 23. Gemeinschaftstagung der Deutschen Arbeitsgemeinschaft für Akustik, DAGA '97*, 614-615.

Kröger, B.J., Schröder, G. Opgen-Rhein, C. (1995): „A gesture-based dynamic model describing articulatory movement data“, *Journal of the Acoustical Society of America* **98,** 1878-1889.

Löfqvist, A., Gracco, V.L. (1999): “Interarticulator programming in VCV sequences: lip and tongue movements”, *Journal of the Acoustical Society of America* **105,** 1864-1876.

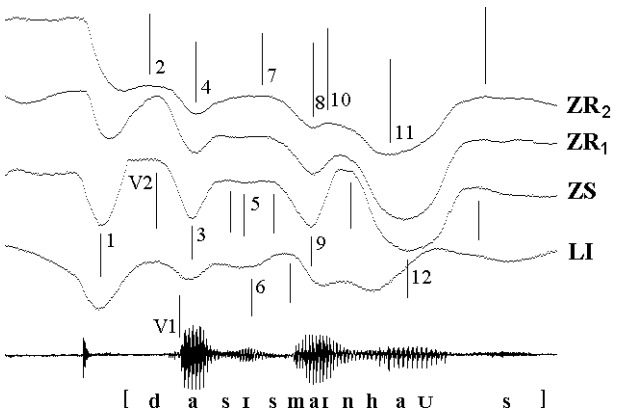

Abbildung 1 Oben: Die gemessenen Artikulatorbewegungen für die Äußerung „Das ist mein Haus“ (Sprecher CS). Auslenkungskurven für mittleren (ZR_2) und vorderen Zungenrücken (ZR_1), für Zungenspitze (ZS) und Unterlippe (LI) jeweils in der artikulatorischen Hauptbewegungsrichtung. Die Zahlen kennzeichnen die artikulatorischen Gesten (vgl. Tab. 1). Die Marken geben Beginn und Ende einer Geste an. Unten: Oszillogramm des zugehörigen akustischen Signals und seine phonetische Transkription.

Tabelle 1 Liste der in der Äußerung „Das ist mein Haus“ (Sprecher CS, Abb. 1) auftretenden artikulatorischen Gesten.

(1)	**veap1**:	apikale Verschlußgeste des [d] in „das“
(2)	**audl1**:	dorsal-labiale Geste des [a] in „das“
(3)	**bvap1**:	apikale Beinahverschlußgeste des [s] in „das“
(4)	**iudl1**:	dorsal-labiale Geste des [ɪ] in „ist“
(5)	**bvap2**:	apikale Beinahverschlußgeste des [s] in “ist“
(6)	**vela1**:	labiale Verschlußgeste des [m] in „mein“
(7)	**audl2**:	dorsal-labiale Öffnungsgeste des [aɪ] in „mein“
(8)	**iudl2**:	dorsal-labiale Schließgeste des [aɪ] in „mein“
(9)	**veap2**:	apikale Verschlußgeste des [n] in „mein“
(10)	**audl3**:	dorsal-labiale Öffnungsgeste des [aʊ] in „Haus“
(11)	**uudl1**:	dorsal-labiale Schließgeste des [aʊ] in „Haus“
(12)	**bvap3**:	apikale Beinahverschlußgeste des [s] in „Haus“