Untersuchung der Faktoren Dauer und Koartikulation bei der Modellierung von Sprechgeschwindigkeit in der Spracherkennung [1]

Britta Wrede, Gernot A. Fink, Gerhard Sagerer
[bwrede|gernot|sagerer]@techfak.uni-bielefeld.de
Technische Fakultät, Universität Bielefeld

Zusammenfassung

Die Variabilität spontan gesprochener Sprache führt zu deutlichen Verminderungen der Leistung von Spracherkennungssystemen. Dabei ist die Sprechgeschwindigkeit ein einflußreicher Variabilitätsfaktor, der sich sowohl in der zeitlichen Dimension Dauer als auch in den spektralen Eigenschaften niederschlägt. Die meisten Ansätze zur Modellierung von Sprechgeschwindigkeit in der Spracherkennung verwenden dauerbasierte Maße. Dauer ist aber nicht der beste Prädiktor für akustische Reduktion. Ein direktes Maß der Reduktion, z.B. mittels Formanten, sagt spetrale Verzerrungen verläßlicher voraus.

Es werden Erkennungsexperimente vorgestellt, die zeigen, daß die Wahl des Kriteriums zum Training dauer- oder reduktionsspezifischer Modelle zwar zu keinen signifikanten Unterschieden in der Erkennungsleistung führt. Da die Maße aber nicht miteinander korreliert sind, liegt die Annahme nahe, daß die verschiedenen Kriterien unterschiedliche Eigenschaften messen und adaptieren.

1 Einleitung

Die Variabilität vor allem spontan gesprochener Sprache stellt ein grundlegendes Problem der Spracherkennung dar. Als eine besonders auffällige Dimension von Variation sticht dabei die Sprechgeschwindigkeit hervor (vgl. z.B. [9]). Dabei stellt sich die Frage, welche Eigenschaften schnell gesprochener Sprache besonders für den Performanzverlust von Spracherkennungssystemen verantwortlich sind.

Ein offensichtliches Merkmal ist die Dauer, deren Variabilität mithilfe der Übergangswahrscheinlichkeiten der Hidden-Markov-Modelle (HMMs) modelliert wird. Dabei restringiert die in solchen Systemen verwendete HMM-Topologie die Dauervariation häufig auf eine gewisse Mindestlänge, die für schnelle Sprache nicht angemessen ist. Aussprachevarianten, die als Folge von Dauerverkürzungen unter anderem durch Elisionen und daraus resultierender Assimilation entstehen, stellen ein weiteres Problem dar, da die Modellierung von Aussprachevarianten zu einer erheblichen Vergrößerung des Suchraumes während der Erkennung führt.

Einer weniger expliziten Untersuchung ist bisher der Einfluß der Veränderung der spektralen Eigenschaften, die mit der Variabilität der Sprechgeschwindigkeit einhergehen, unterzogen worden. Dabei zeigen phonetische Untersuchungen, daß u.a. systematische Veränderungen der Formantenfrequenzen stattfinden.

[1] Diese Arbeit wurde teilweise im Rahmen des Graduiertenkollegs “Aufgabenorientierte Kommunikation” an der Universität Bielefeld von der Deutschen Forschungsgemeinschaft gefördert.

In Proc. Konferenz Elektronische Sprachsignal-Verarbeitung, Bonn, 2001.

Diese Systematik wird in den Ansätzen zur Modellierung von Sprechgeschwindigkeit jedoch nicht berücksichtigt. Ein weit verbreiteter Ansatz aus der Spracherkennung ist die Verwendung tempospezifischer Modelle, die unabhängig voneinander trainiert werden ([6, 12, 8]). Dazu wird eine Schätzung der Sprechgeschwindigkeit verwendet, deren Maß in der Regel dauerbasiert ist.

Phonetische Untersuchungen haben jedoch gezeigt, daß Dauer nicht immer ein guter Prädiktor für Veränderungen der Formantenfrequenzen ist. So konnte gezeigt werden, daß der Sprechstil Einfluß darauf nimmt, ob die Artikulation mit zunehmender Sprachgeschwindigkeit beschleunigt wird, und somit, in welcher Art sich die spektralen Eigenschaften verändern ([7]).

Im folgenden soll daher untersucht werden, inwieweit ein dauerbasiertes Maß bei der Modellierung von Sprechgeschwindigkeit sich von einem formantenbasierten Maß der spektralen Veränderungen unterscheidet.

2 Geschwindigkeit und Koartikulation

Ein systematischer Zusammenhang zwischen Sprechgeschwindigkeit und spektralen Merkmalen ist in vielen Experimenten beobachtet worden. So postulierte Lindblom in [5] Zielfrequenzen von Formanten, sogenannte Targets, die aus den tatsächlich beobachteten Formantenfrequenzen in Abhängigkeit von ihrer Dauer und ihrem Kontext interpoliert werden. In realisierten Vokalen werden diese Targets jedoch mit zunehmender Sprechgeschwindigkeit nicht mehr erreicht, so daß ein *Target Undershoot* entsteht.

In weiteren Experimenten konnte ein Zusammenhang zwischen Sprechgeschwindigkeit und Reduktion festgestellt werden, wobei jedoch auch andere Faktoren wie Dauer ([4, 10, 13]) und Betonung ([2, 4, 10]), die zum Teil unabhängig von Sprechgeschwindigkeit sind, untersucht wurden.

Je nach Sprechstil können sich diese Faktoren unterschiedlich stark auf die Reduktion auswirken. So konnte in [7] gezeigt werden, daß bei deutlichem Sprechstil das Phänomen des Target-Undershoot vermieden wird, indem die Artikulatoren schnellere Bewegungen ausführen. Diese Strategie spiegelt sich in steileren Formantenbewegungen wider.

Ein einheitliches Maß für Reduktion und Koartikulation gibt es in der Literatur nicht. Da sich diese Phänomene besonders deutlich auf die Formantenfrequenzen auswirken, werden in der Regel formantenbasierte Maße verwendet. In [1] werden für Zentralisierung und Koartikulation zwei unterschiedliche Maße als Indikatoren für die Güte der Artikulation berechnet. Als Maß für die Zentralisierung wird die euklidische Distanz der ersten beiden Formanten der beobachteten Vokale zum Mittelpunkt des Vokaldreiecks bestimmt [1, S. 107f]. Dazu wurden die barkskalierten Frequenzen mittelwertbereinigt und varianznormiert. Je geringer die Distanz der Formanten zum Zentrum, desto stärker gilt der Vokal als reduziert.

Für die Quantifizierung der Koartikulation wurde in [1, S. 109ff] ein Target-Undershoot-Maß definiert, das auf der Berechnung von Dichten im Formantenraum beruht. Diese Dichten werden aus besonders deutlich gesprochenen Vokalen berechnet und als Targets interpretiert. Die Wahrscheinlichkeit, mit der ein beobachteter Vokal einer solchen Wahrscheinlichkeitsverteilung zugeordnet wird, gibt den Grad der Koartikulation an. Je geringer diese Wahrscheinlichkeit ist, desto weiter entfernt liegt der Vokal zu einer der Target-Verteilungen und als um so stärker koartikuliert gilt er.

In [11] wird zur Bestimmung der Koartikulation von Konsonanten die Neigung der F2-Verläufe betrachtet. Die zugrundeliegende Annahme dabei ist, daß ein Konsonant um so stärker koartikulatorischen Prozessen unterworfen ist, je weiter die Artikulationsorte der benachbarten Laute entfernt liegen. Diese Entfernung, die die Artikulatoren zurücklegen müssen, spiegelt sich besonders in dem Verlauf des zweiten Formanten wider. Die Differenz der Steigungen dieses Formanten, gemessen am Ende und zu Beginn eines Konsonanten, soll diesen Einfluß quantifizieren.

In der gleichen Untersuchung wird ein weiteres Merkmal als Unterstützung der Segmentierung verwendet, das die Neigung des Spektrums mißt. Im Frequenzschwerpunkt (*Center of Gravity* COG), der als gewichteter Mittelwert der Frequenzbänder berechnet wird, spiegelt sich diese Neigung wider. Je niedriger der Wert des COG, desto weniger stark ist das Spektrum geneigt. In [11] wird argumentiert, daß sich Reduktion in tieferen COG Werten niederschlägt.

3 Sprechgeschwindigkeit in der Spracherkennung

In der Spracherkennung finden solche systematischen Veränderungen der spektralen Eigenschaften gesprochener Sprache sowie deren Messung kaum Berücksichtigung. Ein häufig gewählter Ansatz zur Modellierung von Sprechgeschwindigkeit besteht hier in dem Training tempospezifischer Modelle. In [6] wird die Trainingsstichprobe anhand der mittleren Sprechgeschwindigkeit pro Äußerung in drei Klassen unterteilt, mit denen dann unabhängig voneinander tempospezifische Modelle trainiert werden. Der Erkennungsphase wird ein ausschließlich auf den dynamischen Merkmalen beruhender Gauß-Klassifikator vorgeschaltet, der die zu erkennende Äußerung in schnell, mittel oder langsam klassifiziert. Aufgrund dieser Klassifikation werden die entsprechenden Modelle für die folgende Erkennung ausgwählt. Mit diesem Ansatz konnte die Wortfehlerrate (WFR) für langsame Sprache um 64% und für schnelle Sprache um 19% reduziert werden.

In [12] wurde ein Multi-Layer-Perzeptron (MLP) trainiert, um Phonemgrenzen im Sprachsignal zu finden, anhand derer die Sprechgeschwindigkeit für die Auswahl tempospezifischer Modelle geschätzt wurde. Dieses Verfahren reduzierte die WFR jedoch nur um 4%.

Für die Schätzung der Sprechgeschwindigkeit wird in [8] das Signal mit 16 Hz tiefpassgefiltert, so daß man im Wesentlichen die Energiemodulation erhält, die als Silbenrate interpretiert wird. Durch die Anspassung der Übergangswahrscheinlichkeiten konnte damit eine Verringerung der WFR um 10% erzielt werden.

Als Maß für die Einteilung des Test- und Trainingsmaterials in die entsprechenden Geschwindigkeitsklassen wird in allen Verfahren eine Phonem- oder Silbenrate verwendet. Lediglich in [6] wird ein Klassifikator eingesetzt, der die spektralen Merkmale in Form der dynamischen Cepstral-Koeffizienten (MFCC) nutzt, um die Testäußerungen in schnell, mittel oder langsam zu klassifizieren.

4 Experimente

Zur Untersuchung des Einflusses von Dauer und Koartikulation auf die Modellierung von Sprechgeschwindigkeit wurde eine Reihe von Adaptionsexperimenten durchgeführt. Dabei wurden ähnlich wie in den vorgestellten Modellierungsansätzen geschwindigkeitsspezifische Modelle trainiert. Allerdings wurden für die Einteilung des Trainingskorpus zusätzlich zur Dauer Maße zur Bestimmung der akustischen Reduktion verwendet.

4.1 Korpus

Das Test- und Trainingsmaterial für die Experimente ist dem Bielefelder SLACC-Korpus (Spoken LAnguage Car Control) entnommen. Es besteht aus gelesenen Anweisungen zur Steuerung nicht-sicherheitsrelevanter Funktionen im Auto. Die Anweisungen wurden vom Beifahrer während der Fahrt gelesen. Die durchschnittliche Äußerungslänge ist mit 5,6 Wörtern sehr gering, was die Einteilung ganzer Äußerungen in Tempoklassen für das Training tempospezifischer Modelle ermöglicht. Insgesamt wurden über 9000 Äußerungen von 5 Sprecherinnen und 12 Sprechern für das Training verwendet, was einer Signaldauer von ca. 560 Minuten entspricht.

4.2 Berechnung der verwendeten Maße

Alle Maße basieren auf einer Segmentierung des Trainingskorpus, die mittels eines forcierten Alignments vorgenommen wurde. Da die Einteilung auf Äußerungsebene erfolgte, wurden die Mittelwerte über jeweils einen Turn berechnet. Für die Vokaldauer (VDUR) wurde demnach die Dauer aller in einer Äußerung vorkommenden Vokale gemittelt. Die Formanten für die Berechnung der formantenbasierten Maße wurden mithilfe des ESPS-Formantentrackers [2] berechnet.

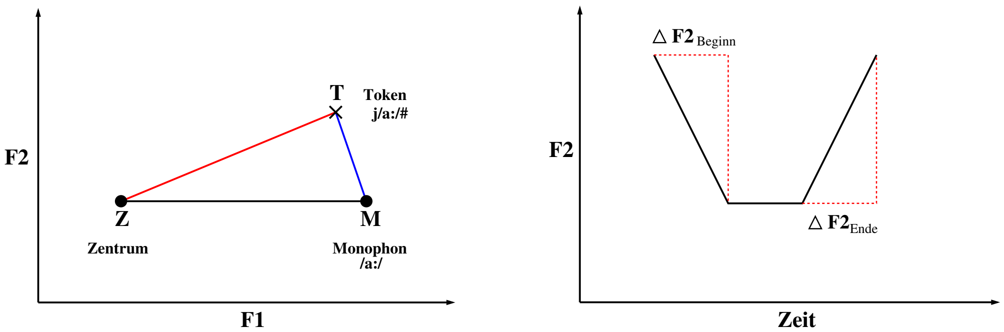

Abbildung 1: Berechnung der formantenbasierten Maße BRED und BCOA sowie der Neigungsdifferenzen des 2. Formanten F2SD

Für das Reduktionsmaß (BRED) wurde im barkskalierten Frequenzraum die euklidische Distanz der ersten beiden Formanten F1 und F2 jedes vorkommenden Vokaltokens T zum Zentrum Z des sprecherspezifischen Vokaldreiecks berechnet und mit der Distanz der mittleren Formantenwerte über alle Token dieses Vokals M zum Zentrum normiert. Als Koartikulationsmaß (BCOA) wurde die gleichermaßen normierte Distanz des Token T zum Monophonmittelpunkt M berechnet (vgl. Abb. 1). Damit ergeben sich folgende Definitionen:

$$\text{BRED} = \frac{||Z - T||}{||Z - M||} \qquad \text{BCOA} = \frac{||M - T||}{||Z - M||}$$

Die Maße BRED und BCOA wurden in Anlehnung an die Reduktions- und Koartikulationsmaße in [1] definiert. Dabei wurden jedoch zwei Vereinfachungen in Kauf genommen: zum einen wurden statt optimaler Targetwerte die Mittelwerte der vorhandenen Token verwendet, da

[2] Das Entropic Signal Processing System (ESPS) ist ein kommerzielles Software Paket von Entropic Research Laboratory, Inc.

keine ”optimalen” Realisierungen von Vokalen vorlagen; zum anderen wurde für das Koartikulationsmaß ebenfalls die euklidische Distanz berechnet, um so die aufwendige Berechnung von Dichten zu vermeiden. Zur Berechnung der Neigungsdifferenz des 2. Formanten (F2SD), die die Koartikulation von Konsonanten messen soll, wurde die 1. Ableitung zu Beginn und Ende der Konsonanten berechnet (Abb. 1). Damit ergibt sich für die Berechnung gemäß [11]

$$\text{F2SD} = \Delta F2_{Ende} - \Delta F2_{Beginn}$$

Neben diesen formantenbasierten Maßen wurde ein weiteres Maß berechnet, das auf die relativ aufwendige Formantenberechnung verzichtet. In Anlehnung an [11] wurde dazu der Frequenzschwerpunkt des Spektrums (COG) gemäß

$$\text{COG} = \frac{\sum_i f_i \cdot E_i}{\sum_i E_i}$$

bestimmt, wobei f_i die Frequenz in Hz und E_i die spektrale Energie dieses Frequenzbandes angibt.

4.3 Systemgrundlagen

Die Experimente wurden mit dem Spracherkennungssystem ESMERALDA [3] durchgeführt. Für die akustische Modellierung der Triphone wurden semi-kontinuierliche HMMs mit 512 Mischverteilungen verwendet. Es wurde alle 10 ms in einem 16 ms Hamming-Fenster ein Merkmalsvektor bestehend aus den ersten 12 Cepstral-Koeffizienten (MFCC) und der Signalenergie sowie deren ersten beiden Ableitungen berechnet. Es wurde kein Sprachmodell verwendet. Das Lexikon bestand aus 658 Wörtern.

Training: Mit diesem System wurden zunächst allgemeine Modelle bis zu ihrem Leistungsmaximum trainiert. Diese allgemeinen Modelle wurden dann mit den entsprechenden Subklassen der Trainingsstichprobe auf spezifische Dauer- bzw. Reduktionsklassen adaptiert.

Erkennung: In der Erkennung wurde keine Temposchätzung vorgenommen. Stattdessen wurden mit allen drei Modellen Erkennungshypothesen erzeugt, von denen die bestbewertete ausgewählt wurde. Dieses Vorgehen stellte sich als deutlich besser heraus als eine optimale Einteilung der Teststichprobe gemäß ihrer durchschnittlichen Vokaldauer bzw. Reduktionsmaße, die auf einer vorgegebenen Segmentierung berechnet wurden (vgl. [14]).

5 Ergebnisse

Die Erkennungsleistung des Referenzsystems mit allgemein trainierten Modellen lieferte eine Wortfehlerrate von 25.0%. Die speziell trainierten Modelle führten durchgängig zu einer signifikanten Verbesserung dieser Leistung (vgl. Tabelle 1). Dabei treten kaum relevante Unterschiede zwischen der Leistung der dauerspezifisch trainierten Modelle (VDUR) zu den koartikulations- bzw. reduktionsspezifisch trainierten Modellen (BCOA, BRED) auf. Die Einteilungen der Modelle gemäß des spektrumbasierten Merkmals COG sowie des Maßes für die Koartikulation von Konsonanten (F2SD) waren deutlich weniger erfolgreich.

	REF	RAND	VDUR	BCOA	BRED	F2SD	COG
WFR	25.0	24.9	23.4	23.2	23.5	23.9	24.1
$\triangle$WFR (in %)	-	-0.4	-6.4	-7.2	-6.0	-4.4	-3.6

Tabelle 1: Wortfehlerraten der dauer- bzw. reduktionsspezifischen Modelle und relative Veränderungen gegenüber der Fehlerrate des Basissystems

Um zu testen, ob die Verbesserungen der Erkennungsleistung nicht auf Artefakte zurückzuführen sind, die durch das veränderte Erkennungsdesign entstehen, wurde zum Vergleich eine zufällige Einteilung der Äußerungen vorgenommen (RAND). Das Ergebnis zeigt, Verbesserungen des Basissystems nicht aufgrund des veränderten Erkennungsszenarios erzielt werden. Bemerkenswert ist, daß die Modelle VDUR ihre beste Leistung bereits nach einer Adaptionsiteration zeigten, während alle anderen Modelle deutlich mehr Iterationen benötigten, um eine ähnliche Erkennungsleistung zu erreichen.

Ein Vergleich der fünf verwendeten Maße zeigte, daß zwischen diesen Maßen, berechnet als Mittelwerte über je eine Äußerung, so gut wie keine Korrelationen bestanden. Lediglich die beiden formantenbasierten Maße BCOA und BRED wiesen ein Korrelation von über .8 auf. Alle anderen Korrelationen lagen unter .3. Dies deutet daraufhin, daß die entsprechenden Modelle unterschiedliche Eigenschaften adaptieren.

6 Diskussion

In der Erkennungsleistung unterscheiden sich die entsprechend der Vokaldauer trainierten Modelle kaum von den entsprechend formantenbasierter Merkmale trainierten Modelle. Das spektrumbasierte Maß bleibt jedoch deutlich hinter den Leistungen zurück. Aufgrund der geringen Korrelationen zwischen den verwendeten Maßen liegt die Annahme nahe, daß die Vokaldauer andere Variabilitäten des Spektrums erfaßt als die formantenbasierten Maße. Es scheint daher erfolgversprechend, eine Kombination dieser Maße für die Einteilung des Trainingsmaterials zu verwenden.

Zudem sollte eine lokale Variation der Tempo- bzw. Reduktionsklasse sowohl im Training als auch in der Erkennung zugelassen werden, indem beispielsweise eine Klassifikation auf Wortebene anstatt wie bisher auf Äußerungsebene erfolgt.

Als Fernziel ist jedoch eine Aufhebung der Klasseneinteilung anzustreben, indem eine systematische Verschiebung der Merkmale durchgeführt wird. Um eine optimale Adaption zu erlangen, ist es essentiell, daß alle Informationen genutzt werden, die die Veränderungen der spektralen Eigenschaften vorhersagen.

Literatur

[1] M. P. Aylett. *Stochastic suprasegmentals: relationships between redundancy, prosodic structure and care of articulation in spontaneous speech.* PhD thesis, University of Edinburgh, 2000.

[2] P. Delattre. An acoustic and articulatory study of vowel reduction in four languages. *Int. Review of Applied Linguistics in Language Teaching*, 7:295–325, 1969.

[3] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In Václav Matoušek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

[4] M. Fourakis. Tempo, stress, and vowel reduction in american english. *J. Acoustical Society of America*, 90(4):1816–1827, 1991.

[5] B. Lindblom. Spectrographic study of vowel reduction. *J. Acoustical Society of America*, 35(11):1773–1781, 1963.

[6] F. Martínez, D. Tapias, and J. Álvarez. Towards speech rate independence in large vocabulary continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 725–728, 1998.

[7] S. J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in english stressed vowels. *J. Acoustical Society of America*, 96(1):40–55, 1994.

[8] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. European Conf. on Speech Communication and Technology*, pages 2079–2082, 1997.

[9] J. Trouvain, J. Koreman, A. Erriquez, and B. Braun. Articulation rate measures and their relation to phone classification in spontaneous and read german speech. In *Proc. Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, France, August 2001. ISCA. To appear.

[10] D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12:1–23, 1993.

[11] R. J. J. H. van Son and L. C. W. Pols. An acoustic description of consonant reduction. *Speech Communication*, 28:125–140, 1999.

[12] J. P. Verhasselt and J. P. Martens. A fast and reliable rate of speech detector. In *Proc. Int. Conf. on Spoken Language Processing*, pages 2258–2261, 1996.

[13] B. Wrede, G. A. Fink, and G. Sagerer. Influence of duration on static and dynamic properties of German vowels in spontaneous speech. In *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 82–85, 2000.

[14] B. Wrede, G. A. Fink, and G. Sagerer. An investigation of modelling aspects for rate-dependent speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, 2001. To appear.