# Die Abbildung akustischer Parameter auf phonetische Merkmale in der automatischen Spracherkennung

Jacques Koreman, Bistra Andreeva, William J. Barry

Universität des Saarlandes, Institut für Phonetik
Postfach 15 11 50, 66041 Saarbrücken

jkoreman@coli.uni-sb.de

Poster

# Die Abbildung akustischer Parameter auf phonetische Merkmale in der automatischen Spracherkennung

Jacques Koreman, Bistra Andreeva, William J. Barry

In diesem Artikel wird beschrieben, wie die Abbildung von akustischen Parametern auf phonetische Merkmale wie [nasal] und [labial] mit Hilfe eines Kohonennetzes durchgeführt wird und welche Vorteile sie für die Identifikation von Konsonanten in einem automatischen Spracherkennungssystem hat. Dabei wird die große Varianz in den akustischen Parametervektoren auf viel homogenere Vektoren von phonetischen Merkmalen abgebildet. Gleichzeitig wird der Teil der Varianz, der nicht zur Unterscheidung der Laute beiträgt, vernachlässigt. Dies führt zu einer verbesserten Konsonantenidentifikation in der Hidden-Markov-Modellierung. Mögliche phonetische Gründe für die Verbesserung der Identifikationsraten im Vergleich zu einem traditionellen Spracherkennungssystem werden erörtet.

In this article, it is described how a Kohonen network can be used to map acoustic parameters onto phonetic features like [nasal] and [labial], and which advantages this mapping has for the identification of consonants in an automatic speech recognition system. The mapping causes the large variation in the vectors of acoustic parameters to be mapped onto much more homogenous vectors of phonetic features. At the same time, the part of the variability in acoustic signal properties which does not help to distinguish between sounds is disregarded. This leads to better consonant identification rates in hidden Markov modelling. Possible phonetic reasons for the improved consonant identification rates in comparison to a traditonal speech recognition system are proposed.

## 1 Einführung

In der automatischen Spracherkennung führt die große Variabilität bei der Realisierung von Konsonanten zu zahlreichen Verwechslungen, infolgedessen Wörter und Wortsequenzen im gesprochenen Text falsch erkannt werden können. Das Problem wird teilweise durch "Top-Down-Verarbeitung" gelöst, weil die meisten phonbasierten Spracherkennungssysteme zusätzlich ein Lexikon und ein Sprachmodell benutzen (Lee, 1990). Trotzdem besteht insbesondere im Hinblick auf die Erkennung spontansprachlicher Äußerungen eine Notwendigkeit, die phonetische Verarbeitung zu optimieren.

Ziel der vorliegenden Arbeit ist es zu zeigen, wie eine linguistisch orientierte Vorverarbeitung des akustischen Signals die Variation ordnet und zu einer verbesserten phonetischen Dekodierung von Konsonanten führen kann (vgl. Bitar & Espy-Wilson, 1995a,b; Kirchhoff, 1996a,b). Anstelle der üblichen direkten Abbildung der akustischen Eigenschaften (Mel-frequenz Cepstralkoeffizienten (MFCK's), Energie und ihre Delta-Parameter) auf die Phonemebene wird eine Zwischenstufe mit phonetischen Merkmalen eingeschaltet, die im Lautsystem eine potentiell bedeutungsunterscheidende Funktion haben. Die phonetischen Merkmale, wie [nasal] und [labial], basieren auf der Konsonantentabelle des IPA. Es wird weder ein Lexikon noch ein Sprachmodell benutzt.

## 2 Daten

Als Inputdaten für das Experiment wurden vorgelesene Texte aus vier unterschiedlichen Sprachen (Deutsch, Englisch, Italienisch und Niederländisch) aus der Eurom0-Datenbank verwendet. Die Texte (mit einer Dauer von anderthalb bis zwei Minuten) wurden jeweils von zwei männlichen und zwei weiblichen Sprechern vorgelesen. Es wurden mehrere Sprachen benutzt, um die Datenmenge zu vergrößern und damit das automatische Spracherkennungssystem besser trainieren zu können.

Bevor sie als Input für das Konsonantenidentifikationssystem[1] verwendet wurden, wurden die Mikrophonsignale zuerst mit 12 MFCK's, Energie und den davon abgeleiteten Delta-Koeffizienten parametrisiert. Die 26 akustischen Parameter wurden mit dem HTK-Programm HCopy (Young et al., 1995) über ein Hamming-Fenster von 15 ms und mit einer Schrittgröße von 5 ms berechnet. Es wurde ein kurzes Fenster gewählt, damit die spektralen Änderungen nicht zu sehr zeitlich verschmiert werden. Die verwendete Präemphase war 0,97.

## 3 Die Architektur des Konsonantenidentifikationssystems

Das System zur Konsonantenidentifizierung besteht aus zwei Teilen: Der erste Teil bildet die akustischen Parameter auf phonetische Merkmale ab, während der zweite Teil die Konsonanten identifiziert (siehe auch Koreman et al., 1997). Die

[1] Es soll darauf hingewiesen werden, daß für die Erkennung der Konsonanten die gleichen Signale verwendet wurden wie für das Training (deswegen sprechen wir von Identifikation).

Konsonantenidentifikation mit dem unten beschriebenen hybriden System wird verglichen mit den Ergebnissen eines traditionellen automatischen Spracherkennungssystems, in dem die akustischen Vektoren direkt mit Hidden-Markov-Modellen (HMM's) auf Phonemebene modelliert werden. Dieses System wird im Text mit dem Terminus "Baselinesystem" bezeichnet.

Die Abbildung von akustischen Parametern auf phonetische Merkmale findet mit Hilfe eines Kohonennetzes (Dalsgaard, 1992) statt. Dieser Teil des Systems berücksichtigt einerseits die akustische Verschiedenheit der Konsonantenrealisierungen, bildet die große Menge an Realisierungen aber auf eine kleine Menge phonetischer Merkmale ab. Die phonetischen Merkmale grenzen alle natürlichen Klassen innerhalb der vier Sprachen voneinander ab und halten damit alle Bedeutungsunterschiede zwischen den Konsonanten fest.

Im zweiten Teil des Systems werden die Konsonanten identifiziert. Dazu werden mit den phonetischen Merkmalsvektoren 31 HMM's trainiert, d.h. ein HMM für jeden Konsonanten. Dabei werden einfache Links-Rechts-Modelle verwendet, die aus 3 echten Zuständen (plus Anfangs- und Endzustand) bestehen. Jeder Zustand wird mit nur einer multivariaten Wahrscheinlichkeitsverteilung trainiert.

## 4 Ergebnisse und Interpretation

Im hybriden Konsonantenidentifikationssystem werden 41,10% der Konsonanten korrekt identifiziert, gegenüber 15,04% im Baselinesystem, in dem die akustischen Parameter direkt benutzt werden, um HMM's zu trainieren. Die stark verbesserte Konsonantenidentifikation hängt mit der vergrößerten Homogenität der HMM's zusammen, die durch die akustisch-phonetische Abbildung erreicht wird.

Einige Laute werden schon im Baselinesystem gut erkannt. Dabei handelt es sich meistens um sprachspezifische Laute (siehe Tabelle 1), so daß zu erwarten ist, daß ihre akustische Variation geringer ist (z.T. auch durch die kleinere Anzahl der Realisierungen). Die Konsonanten, die trotz höherer Anzahl der Realisierungen gut identifiziert werden, nämlich /ɹ/ und /w/, kommen nur im Englischen bzw. im Englischen und Italienischen vor.

| Konsonant | % korrekt Baseline | % korrekt Hybrid | n | Sprache |
|---|---|---|---|---|
| ç | 98,9 | 59,8 | 87 | Deutsch |
| ð | 78,5 | 63,4 | 93 | Englisch |
| ɲ | 100,0 | 100,0 | 4 | Italienisch |
| ʎ | 100,0 | 100,0 | 12 | Italienisch |
| θ | 100,0 | 93,3 | 20 | Englisch |
| ʒ | 90,0 | 75,0 | 10 | Englisch, Italienisch |
| ɹ | 86,9 | 69,6 | 122 | Englisch |
| w | 82,4 | 71,4 | 153 | Englisch, Italienisch |

Tabelle 1. Konsonanten mit einer Identifikationsrate von mehr als 75% im Baselinesystem (% korrekt = Identifikationsrate; n = Anzahl der Realisierungen des Konsonanten)

In der Verwechslungsmatrix des hybriden Experiments geht der sprachspezifische Vorteil verloren. Der Grund dafür ist wahrscheinlich, daß die redundanten akustischen Eigenschaften nicht mehr in den Inputvektoren zu den HMM's repräsentiert sind, was dazu führt, daß einige von den sprachspezifischen Konsonanten jetzt schlechter identifiziert werden.

Dafür werden aber mit nur wenigen Ausnahmen alle anderen Konsonanten viel besser identifiziert. Wenn wir uns z.B. die Identifikationsrate für /l/ anschauen, stellen wir fest, daß es im Baselinesystem schlecht identifiziert wird (19,6%). Grund dafür ist, daß alle Realisierungen des /l/ in *einem* HMM modelliert werden müssen, obwohl sie zum Teil akustisch sehr unterschiedlich sind. Die akustische Variation bei der Realisierung von /l/ wird in Abbildung 1 gezeigt. In der phonotopischen Organisation des Kohonennetzes gibt es zwei Lokalisierungen für /l/: ein helles /l/ mit meistens alveolaren Lauten in seiner Umgebung, und ein dunkles /l/ mit meistens velaren und uvularen Lauten in seiner Umgebung. Die akustische Variabilität bei den /l/-Realisierungen wird im Kohonennetz auf die gleichen phonetischen Merkmale abgebildet[2], so daß die

[2] Das Problem kann zum Teil dadurch gelöst werden, daß nicht eine, sondern mehrere Wahrscheinlichkeitsverteilungen pro Zustand im HMM verwendet werden, um die Beobachtungen zu modellieren. Dabei wird dem Experimentator überlassen, für jeden Laut die Anzahl der Wahrscheinlichkeitsverteilungen pro Zustand zu bestimmen. Das

Inputparameter für das HMM viel homogener sind. Dadurch wird das /l/ im hybriden System in 53,8% der Fälle korrekt identifiziert. Diese Tendenz beschränkt sich nicht nur auf das /l/.

| | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|
| **26** | alv_ | alv_ | n | n | n | n |
| **27** | lab_ | alv_ | n | w | w | w |
| **28** | n | w | w | _alv | vel_ | _alv |
| **29** | n | w | l | l | lab_ | m |
| **30** | n | n | l | l | _alv | _lab |
| **31** | n | n | n | l | _lab | m |
| **32** | n | n | n | n | n | n |
| **33** | n | n | n | n | n | n |

| | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|
| **25** | R | R | R | x | f | s |
| **26** | R | R | _alv | x | x | x |
| **27** | l | _alv | x | _vel | x | x |
| **28** | l | l | x | _alv | l | x |
| **29** | l | l | l | l | l | x |
| **30** | l | l | x | _alv | f/p0 | p0 |
| **31** | p0 | x | x | x | p0 | p0 |
| **32** | k | x | x | s | s | f |

Abb. 1. Zwei Ausschnitte aus dem Kohonennetz mit meist alveolaren bzw. velaren und uvularen Lauten[3] (die Numerierung zeigt die Position der Reihen und Spalten im Kohonennetz)

Aufgrund des beschränkten Platzes, der uns zur Verfügung steht, werden weitere unterstützende Daten auf der Tagung präsentiert.

## 5 Schlußfolgerungen

In diesem Artikel haben wir gezeigt, daß Konsonanten durch die Abbildung akustischer Parameter auf phonetische Merkmale besser identifiziert werden. Diese Verbesserung wird dadurch erreicht, daß die HMM's mit homogeneren Inputvektoren trainiert werden. Die größere Homogenität entsteht durch die "Vernachlässigung" von nicht-distinktiven Signaleigenschaften. In dem

---

Kohonennetz nimmt einem die Qual dieser Wahl größtenteils ab, da es innerhalb seiner Dimensionen das gesamte akustische Material optimal verteilt.

[3] Folgende sind keine standard IPA-Symbole: p0 = simmloser Verschluß; _lab / lab_ = Vokaltransition zu bzw. von einem labialen Laut; _alv / alv_ = Vokaltransition zu bzw. von einem alveolaren Laut; _vel / vel_ = Vokaltransition zu bzw. von einem velaren Laut. Die Vokaltransitionen, die hier nicht für die Erkennung verwendet wurden, sind vorhanden, weil das Kohonennetz aus einem anderen Experiment übernommen wurde (siehe Koreman et al., 1997).

sprachübergreifenden Experiment, das in der vorliegenden Arbeit vorgestellt wurde, führt die Vernachlässigung dieser Signaleigenschaften einerseits dazu, daß manche sprachspezifische Konsonanten etwas schlechter identifiziert werden, gleichzeitig aber führt sie zur Verbesserung der Identifikationsrate der meisten nicht-sprachspezifischen Konsonanten.

## 6 Literaturangaben

Bitar, N. & Espy-Wilson, C. (1995a). Speech parameterization based on phonetic features: application to speech recognition. *Proc. 4th European Conference on Speech Communication and Technology,* 1411-1414.

Bitar, N. & Espy-Wilson, C. (1995b). A signal representation of speech based on phonetic features. *Proc. 5th Annual Dual-Use Techn. and Applications Conf.,* 310-315.

Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and Gaussian probability density functions. *Computer Speech and Language* **6**, 303-329.

Kirchhoff, K. (1996a). Syllable-level desynchronisation of phonetic features for speech recognition. *Proc. Int. Conf. on Spoken Lang. Proc.,* 2274-2276.

Kirchhoff, K. (1996b). Phonologisch strukturierte HMMs zur automatischen Spracherkennung. In: D. Gibbon (ed.). *Natural Language Processing and Speech Technology (Proceedings of the 3d KONVENS Conference),* 55-63. Berlin: Mouton de Gruyter.

Koreman, J., Barry, W.J. & Andreeva, B. (1997). Relational phonetic features for consonant identification in a hybrid ASR system. *PHONUS* **3**, 83-109. Saarbrücken: Institut für Phonetik, Universität des Saarlandes.

Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transitions on Acoustics, Speech and Signal Processings,* 599-609.

Young, S., Jansen, J., Odell, J., Ollason, D. & Woodland, P. (1995). *The HTK Book.* Cambridge: Cambridge University.