Gehörgerechte Vorverarbeitung für die Spracherkennung auf Basis von Wortuntereinheiten

Christine Hartmann*, Michael Kleinschmidt, Jürgen Tchorz und Birger Kollmeier AG Medizinische Physik,
Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
*christin@medi.physik.uni-oldenburg.de Webseite: http://medi.uni-oldenburg.de/members/juergen/asr.html

I. Einleitung

Bei der Spracherkennung steigt der Trainings- und Erkennungsaufwand überproportional mit dem zu erkennenden Wortschatz an. Um diesem entgegenzuwirken, sucht man kurze Wortuntereinheiten, mit denen sich hohe Erkennungsraten erzielen lassen. In diesem Beitrag wurde die Vorverarbeitung durch das Perzeptionsmodell (PEMO) nach Dau et al. [dau96], welches sich bei der Einzelworterkennung von Ziffern besonders im Störgeräusch bewährt hat [tch99], auf seine Anwendbarkeit bei der Phonemerkennung untersucht. Bei dem Übergang von Einzelworterkennung zur Phonemerkennung nehmen Koartikulationseffekte an Bedeutung zu.
Aufgrund der sehr viel kürzeren zeitlichen Dauer eines Phonems (ca. 30 bis 100 ms) gegenüber einem Wort (ca. 200 bis 500 ms) wurde hier untersucht, inwieweit die unterschiedlichen Komponenten der Merkmalsextraktion sich aufgrund ihrer teilweise langen Zeitkonstanten auf die Klassifikationsleistung auswirken.

II. Vorverarbeitung mit PEMO

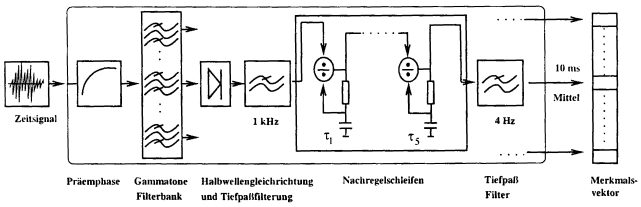

Abbildung 1: PEMO-Verarbeitungsstufen, Quelle: [kle98]

Das ursprünglich für die Modellierung psychoakustischer Experimente entwickelte Perzeptionsmodell (PEMO, Abb. 1) besteht aus mehreren Verarbeitungsstufen. Die fünf Nachregelschleifen und der abschließende Tiefpaß von PEMO sind von großer Relevanz für die Übertragbarkeit von Einzelworterkennung auf Phonemerkennung. In früheren Arbeiten von Tchorz et al. [tch99] haben sich bei der Einzelworterkennung die psychoakustisch ermittelten Zeitkonstanten (5ms, 50ms, 129ms 250ms und 500ms) für die Nachregelschleifen als optimal erwiesen, sowie ein abschließender Tiefpaß von 4 Hz.

III. Zeitkonstanteneffekte

In Abbildung 2 ist die "interne Repräsentation" des Wortes "*Zugbegleiter*" nach Verarbeitung mit PEMO dargestellt. Abbildung 3 zeigt die interne Repräsentation des gleichen Wortes, nur wurde dabei jedes Phonem aus seinem Kontext herausgeschnitten, separat mit PEMO verarbeitet und anschließend wurden die internen Repräsentationen wieder zusammengefügt. Die vertikalen Linien markieren die Phonemgrenzen entsprechend der Labelung der Sprachdatenbank PhonDat. Die internen Repräsentationen wurden mit den Standardparametern wie oben aufgeführt berechnet.
Wie in den Abbildungen 2 und 3 veranschaulicht, beeinflussen zeitlich vorangehende Phoneme in einem Signal die interne Repräsentation nachfolgender Phoneme. Die dargestellte interne Repräsentation ist in Abbildung 2 kontrastreicher als Abbildung 3. Diese starken Kontraste sind bedingt durch Ein- und Ausschwingvorgänge in den Nachregelschleifen. Werden diese Einflüsse von vorhergehenden Phonemen durch die einzelne Verarbeitung der Phoneme eliminiert, so ist die Intensität der internen Repräsentation jedes Phonems größer (Abbildung 3).

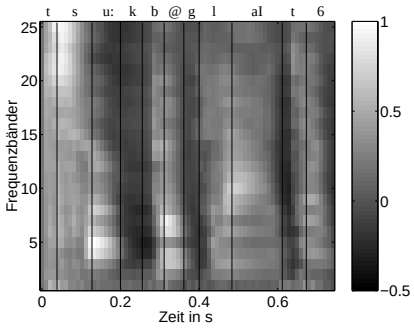

Abbildung 2: Interne Repräsentationen des Wortes *Zugbegleiter*

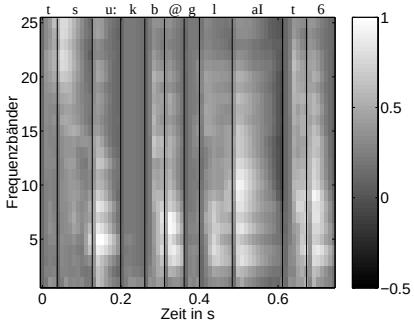

Abbildung 3: Interne Repräsentationen der Phoneme $t,s,u:,k,b,@,g,l,aI,t,6$

IV. Spracherkennungsexperimente

IV.A Experimenteller Rahmen

Für die Spracherkennungsexperimente wurde ein HMM-Erkenner (HTK) mit 3 emittierenden Zuständen pro Phonem und kontinuierlicher Ausgabedichteverteilung benutzt. Als Sprachmaterial wurden insgesamt 2908 Geschichten und Sätze der deutschsprachigen Datenbank PhonDat verwendet. Unter Berücksichtigung der Phonemhäufigkeit, Sprecher, Sprechergeschlecht und Aufnahmeorte (Kiel, Bochum, München) wurde ein Trainings- und Testkorpus wie folgt ausgewählt.
Trainingskorpus:

- Anzahl Sprecher: weiblich: 33, männlich: 29
- Anzahl der Sätze und Geschichten: 2483
- Anzahl der Phoneme: 93691

Testkorpus:

- Anzahl Sprecher: weiblich: 8, männlich: 10
- Anzahl der Sätze und Geschichten: 425
- Anzahl der Phoneme: 18498

Trainings- und Testkorpus sind disjunkt in Bezug auf die Sprecher. Die Anzahl der für die Erkennung benutzten Phoneme beträgt 41. Die Sprachdaten wurden mit 48kHz Abtastfrequenz aufgenommen und mit 16kHz abgespeichert.

IV.B Experimente

Die Sprachdaten wurden mit dem Perzeptionsmodell vorverarbeitet. Zur besseren Anpassung an den HMM-Erkenner wurden

die Merkmalsvektoren nach [kas99] mit einer Cosinustransformation in eine cepstrale Darstellung mit Hilfe des Programms PECE umgewandelt. Während bei den Trainingsdaten die zeitliche Abfolge der Phoneme bekannt war, wurden die Testdaten **ohne** Phonemsegmentierung vorgegeben.
In mehreren Experimenten wurden die Parameter für die zeitliche Verarbeitung des Perzeptionsmodells (Zeitkonstanten der Nachregelschleifen, abschließender Tiefpaß) variiert, um ihren Einfluß auf die Phonemerkennungsrate (PER) zu untersuchen. Im folgendem werden die Zeitkonstanten der Nachregelschleifen als $\tau_{1,2,3,4,5}$ notiert, wobei τ_i die Zeitkonstante der i-ten Nachregelschleife ist. Die benutzte Gammatone Filterbank hatte 20 Filter und es wurde keine Präemphase verwendet, was sich in Vorversuchen als günstig erwiesen hatte.
Dem Merkmalsvektor wurde die Energie als weitere Dimension hinzugefügt, aber keine Δ oder $\Delta\Delta$-Merkmale.
Die Erkennungsrate wurde berechnet nach:

$$C = \frac{\text{Anzahl der richtig erkannten Phoneme}}{\text{Anzahl der gesamten Phoneme}} \cdot 100\%$$

Als Referenz für die Erkennungsleistung diente eine Mel-Cepstral Vorverarbeitung. Für die Mel-Cepstral Vorverarbeitung wurden 26 Filterbankkanäle, 12 Cepstren und die Energie als Merkmalsvektor benutzt. Zusätzlich wurden die dynamischen Merkmale Δ und $\Delta\Delta$ des Merkmalsvektors berechnet.

V. Ergebnisse

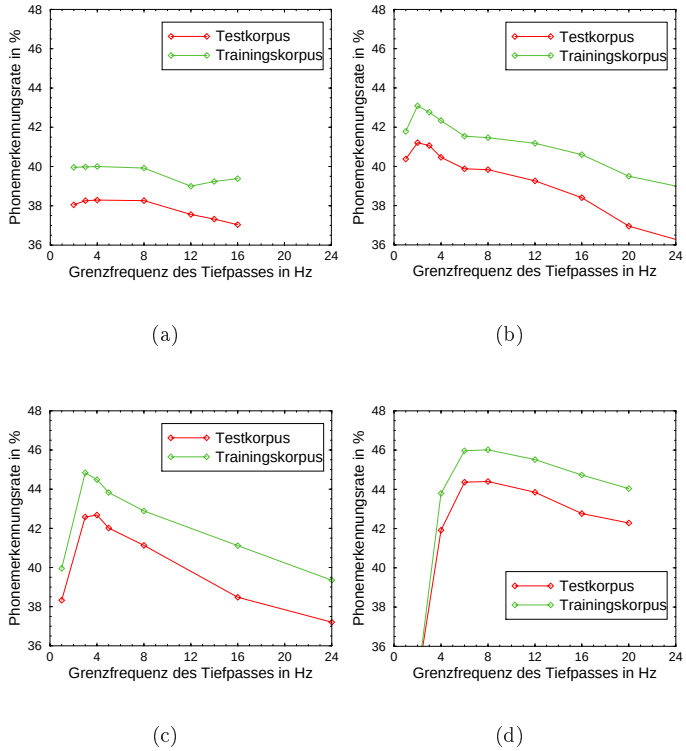

Abbildung 4: Phonemerkennungsraten (PER) in Abhängigkeit von der Grenzfrequenz des abschließenden Tiefpasses für verschiedene $\tau_{1,2,3,4,5}$

Benutzte Einstellungen:
4a) PER für $\tau_{1,2,3,4,5}$=(5 ms, 50 ms, 129 ms, 253 ms, 500 ms)
4b) PER für $\tau_{1,2,3,4,5}$=(5 ms, 50 ms, 129 ms, 253 ms, 500 ms)*0.5
4c) PER für $\tau_{1,2,3,4,5}$=(5 ms, 50 ms, 129 ms, 253 ms, 500 ms)*0.3
4d) PER für $\tau_{1,2,3,4,5}$=(5 ms, 50 ms, 129 ms, 253 ms, 500 ms)*0.1

Bei einer Verkleinerung der Zeitkonstanten der Nachregelschleifen verschiebt sich die PER-Kurve zu höheren Grenzfrequenzen des Tiefpasses (4a bis 4d). Die beste Erkennungsleistung (44,4%) wird erzielt mit $\tau_{1,2,3,4,5}$=(0.5 ms, 5 ms, 12.9 ms, 25.3 ms, 50 ms) und einer Grenzfrequenz von 6 bis 8 Hz. Damit ist PEMO der Mel-Cepstral Vorverarbeitung unterlegen, deren PER bei 51,4% liegt. Abbildung 5 zeigt die normierten (Quasi-) Modulations-Übertragungsfunktionen (MTF) von PEMO für die Abschneidefrequenzen der Tiefpässe f_G, bei denen die Maxima der PER für die jeweiligen Zeitkonstanten $\tau_{1,2,3,4,5}$ aus den Abbildungen 3b, 3c und 3d erreicht wurden. Diese Übertragungsfunktionen zeigen das Modulationsübertragungsverhalten der Nachregelschleifen und des Tiefpasses von einer sinusförmig modulierten Einhüllenden.

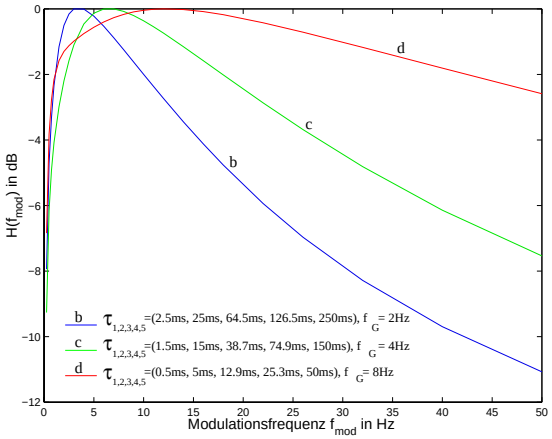

Abbildung 5: (Quasi-) Modulations-Übertragungsfunktion von PEMO für verschiedene Zeitkonstanten der Nachregelschleifen und des abschließenden Tiefpasses

Aus Abbildung 5 ist ersichtlich, daß das Maximum der MTF mit Verkleinerung der Zeitkonstanten der Nachregelschleifen von der Modulationsfrequenz $f_{mod} = 3.5$ Hz über $f_{mod} = 7$ Hz zu $f_{mod} = 12$ Hz wandert.

VI. Schlußfolgerungen und Ausblick

Es konnte gezeigt werden, daß sich PEMO als Vorverarbeitung für Spracherkennung auf Basis von Phonemen prinzipiell eignet, aber geringere Erkennungsraten als Standard Mel-Cepstral-Koeffizienten erreicht.
Die beste Erkennungsleistung wurde mit Zeitkonstanten erzielt, bei denen die Modulationsübertragungsfunktion von PEMO ein Maximum bei 12 Hz hat. Bei dieser Modulationsübertragungsfunktion sind die Phoneme in der internen Repräsentation am deutlichsten ausgeprägt. Den Modulationsfrequenzen um 12 Hz kommt bei der Phonemerkennung eine große Bedeutung zu im Gegensatz zur Einzelworterkennung, bei der gezeigt wurde, daß ein Modulationsübertragungsmaximum um 4 Hz optimal ist.
Die relativ zur Mel-Cepstral Vorverarbeitung schlechte Phonemerkennungsrate von PEMO deutet darauf hin, daß trotz verkleinerter Zeitkonstanten bei der Merkmalsextraktion von PEMO das vorhergehende Signal die interne Repräsentation eines Phonems beeinflußt. In weiteren Studien wird untersucht werden, inwieweit diese Eigenschaft von PEMO bei der Spracherkennung auf Basis von Diphonen und Triphonen nützlich ist.

Literatur

[dau96] Dau, T., Püschel, D., and Kohlrausch, A., "A quantitative model of the "effective" signal processing in the auditory system. Model structure "J. Acoust. Soc. Am., pp. 3615-3622, 1999

[kas99] Kasper, K., Reiniger, H., "Evaluation of PEMO in roust speech recognition" J. Acoust. Soc. Am., Proceedings, p. 1157, 1999

[kle98] Kleinschmidt, M., Tchorz, J., Wittkop, T., Hohmann, V.und Kollmeier, B. "Robuste Spracherkennung durch binaural Richtungsfilterung und gehörgerechte Vorverabeitung", "Fortschritte der Akustik - DAGA 1998", Zürich, pp. 396-397, DEGA, Oldenburg

[tch99] Tchorz, J. and Kollmeier, B., "A model of auditory peception as front end for automatic speech recognition" J. Acoust. Soc. Am. 106 (4), p. 2040-2050, 1999

Besonderen Dank an Klaus Kasper und Herbert Reiniger von der Universität Frankfurt, insbesondere für die PECE Implementation.
Unterstützt von der DFG.