Perzeptive Vorverarbeitung und automatische Selektion sekundärer Merkmale zur robusten Spracherkennung

Michael Kleinschmidt und Volker Hohmann
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
michael@medi.physik.uni-oldenburg.de http://medi.uni-oldenburg.de/members/michael

I. Einleitung

Für die automatische Spracherkennung ist die mangelnde Robustheit gegenüber additiven Störgeräuschen eines der ungelösten Probleme. In diesem Beitrag wird die neuartige Kombination einer perzeptiven Vorverarbeitung mit einem speziellen neuronalen Netz vorgestellt und ihre Robustheit gegenüber Störschallen evaluiert. Das Perzeptionsmodell (PEMO) nach Dau et al. [dau96] ist ein effektives Modell der auditorischen Signalverarbeitung und wurde ursprünglich zur Simulation psychoakustischer Experimente konzipiert. Es konnte bereits erfolgreich zur Vorverarbeitung bei der robusten Einzelworterkennung im Störgeräusch eingesetzt werden [kle98]. In diesem Beitrag wird die Kombination des PEMO mit dem *Feature Finding Neural Network* nach Gramß [gra92] evaluiert. Dabei werden aus den primären (PEMO) Merkmalen zunächst neue temporale und spektrale Merkmale extrahiert und diese anschließend mit Hilfe eines linearen neuronalen Netzes klassifiziert. Diese sekundären Merkmale werden während der Trainingsphase automatisch optimiert. Die vorgestellten Ergebnisse zur sprecherunabhängigen Klassifikation isolierter deutscher Ziffern in unterschiedlichen Störschallsituationen weisen auf eine Verbesserung der Robustheit im Vergleich zu anderen Erkennungssystemen hin.

II. Spracherkennungssystem

Mit dem vorgestellten Spracherkennungssystem wird die robuste Merkmalsextraktion basierend auf dem Perzeptionsmodell nach Dau et al. [dau96] mit dem *Feature Finding Neural Network* nach Gramß [gra92] kombiniert. Beide Verfahren orientieren sich an Erkenntnissen aus Physiologie und Psychoakustik, sind jedoch bisher noch nicht in dieser Kombination evaluiert worden.

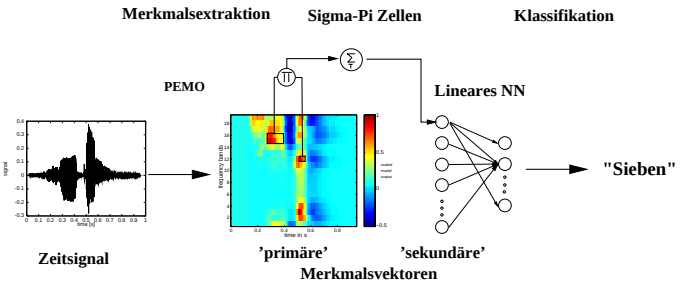


Abb. 1: *Ablauf der Klassifikation mit PEMO/FFNN*

A. Gehörgerechte Vorverarbeitung

Das Perzeptionsmodell (PEMO) nach Dau et al. [dau96] ist ein funktionelles Modell der Signalverarbeitung im peripheren auditorischen System. Es ist in der Lage, das Antwortverhalten von Versuchspersonen in einer Vielzahl von psychoakustischen Experimenten quantitativ nachzubilden. Das PEMO wandelt ein eintreffendes akustisches Signal in die dazugehörige *interne Repräsentation* um, welche sich bereits als ein robustes Merkmal für die automatische Spracherkennung bewährt hat [tch99]. Insbesondere in Kombination mit einem neuronalen Netz als Klassifikator übertrifft die PEMO Vorverarbeitung konventionelle Mel-Cepstralkoeffizienten deutlich an Robustheit gegenüber additiven Störgeräuschen [tch97]. Das verwendete Modell unterzieht die Eingangssignale einer Präemphase und einer Filterbank-Zerlegung mit *Gammatone*-Filtern. Darauf folgen kanalweise Halbwellengleichrichtung und Einhüllendenbildung. Fünf nichtlinear komprimierende Nachregelschleifen und ein Modulationstiefpass bilden Adaptationseffekte nach. Die so erhaltene interne Repräsentation des Zeitsignals wird zur Datenreduktion noch kanalweise über 10 ms gemittelt und bildet damit einen geeigneten Merkmalsvektor. In dieser Anwendung werden 15 auf einer ERB-Skala äquidistant angeordnete Filter mit Mittenfrequenzen zwischen 50 und 7000 Hz verwendet. Im Gegensatz zu früheren Untersuchungen betrug die Filterbreite zwei ERB. Dies entspricht eher den *articulation bands* für die Sprachwahrnehmung des Menschen, welche nach Fletcher und Allan [all94] die doppelte Breite der kritischen Bänder haben.

B. Extraktion sekundärer Merkmale

Von Gramß et al. [gra92] wurde das *Feature Finding Neural Network* (FFNN) als ein Algorithmus zur Spracherkennung vorgeschlagen. Hierbei sollen zunächst aus den vorhandenen (*primären*) Merkmalsvektoren spezielle *sekundäre* Merkmale extrahiert werden, welche dann über ein einstufiges, lineares neuronales Netz klassifiziert werden. Diese sekundären Merkmale werden *Sigma-Pi Zellen* genannt und berechnen sich aus der Folge primärer Merkmalsvektoren $\vec{m}_(t)$ wie folgt:

$$x(f,t,f_0,t_0,\Delta f,\Delta t) = m_f(t) \cdot \sum_{f'=0}^{\Delta f-1} \sum_{t'=0}^{\Delta t-1} m_{f+f_0+f'}(t+t_0+t')$$

Dabei werden ein großes und ein kleines Fenster über die in Matrixform aneinandergereihten primären Merkmalsvektoren gelegt und jeweils die Summe der Werte über ein großes Fenster mit dem Wert eines kleinen Fensters multipliziert. t und f stehen dabei für Zeit- und Frequenzachse der primären Merkmale, während f_0 und t_0 den Abstand der beiden Fenster sowie Δf und Δt die Größe des zweiten Fensters, jeweils in Zeit- und Frequenzrichtung, angeben. Für die Einzelworterkennung werden die so erhaltenen $x(f,t)$ noch über den gesamten Zeitraum summiert, so dass für jedes einzelne Zeitsignal genau ein sekundärer Merkmalsvektor $\vec{x}$ resultiert. Die Form dieser sekundären Merkmale als Sigma-Pi Zellen war von Gramß [gra92] bereits physiologisch und psychoakustisch motiviert worden. Die Ähnlichkeit dieser Sigma-Pi Zellen mit aktuellen Erkenntnissen aus der psychoakustischen Forschung ist jedoch besonders interessant. Insbesondere die mittels der *Reverse Correlation* Methode gewonnenen für die akustische Wahrnehmung von periodischem Rauschen [kae00] wichtigen Merkmale sind in ihrer spektral-temporalen Gestalt und Ausdehnung mit den für die Spracherkennung verwendbaren Sigma-Pi Zellen vergleichbar.

C. Trainingsalgorithmus

Dank der Verwendung eines linearen Klassifikators kann die optimale Gewichtsmatrix analytisch ermittelt werden. Sei N die Zahl der Merkmale, M die Zahl der Klassen und P die Gesamtzahl der zu lernenden Muster (Klassen × Sprecher × Repräsentation pro Klasse und Sprecher) und beinhaltet $\mathbf{X}$ in jeder Spalte den sekundären Merkmalsvektor einer Repräsentation, so lautet das Klassifikationsproblem:

$$\begin{array}{ccccc} \tilde{\mathbf{Y}} & = & \mathbf{W} & \cdot & \mathbf{X} \\ (M \times P) & = & (M \times N) & \cdot & (N \times P) \end{array}$$

Nach [gra92] ergibt sich die optimale Gewichtsmatrix $\mathbf{W}$ zu

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^{+} \text{ , wobei } \mathbf{X}^{+} = \mathbf{X}^{T}\mathbf{\Psi} \text{ , } \mathbf{\Psi} = (\mathbf{X}\mathbf{X}^{T})^{-1}$$

falls der Rang von $\mathbf{X}$ maximal ist und $P \geq N$ gilt. $\mathbf{X}^{+}$ ist die *Pseudoinverse* von $\mathbf{X}$. Da der Trainingsaufwand für gegebene Art und Anzahl von sekundären Merkmalen relativ gering ist, lässt sich die Auswahl von geeigneten sekundären Merkmalen automatisieren. Entsprechend der *Substitutionsregel* nach Gramß [gra92] werden zunächst N zufällige Merkmale verwendet und in jeden Optimierungsschritt das am wenigsten geeignete Merkmal durch ein zufälliges anderes ersetzt. Als Maß für die Eignung wird hier die Relevanz R_i eines Merkmals i verwendet, welche sich aus dem euklidischen Fehlermaß $E = \|\mathbf{Y} - \tilde{\mathbf{W}}\mathbf{X}\|^2$ wie folgt ergibt:

$$R_i = \Delta E_i = E(\text{ohne Merkmal i}) - E(\text{mit Merkmal i})$$

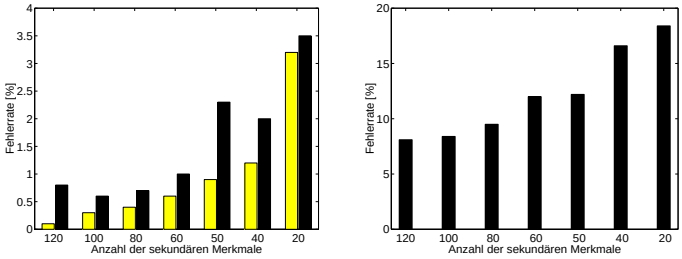
In den folgenden Experimenten wurde jeweils über 500 Iterationen trainiert.

III. Experimente

Die hier vorgestellten Experimente beschränken sich auf isolierte Einzelworterkennung. Als Sprachkorpus fanden Teile des ZIFKOM Sprachdatensatzes der Deutschen Telekom Verwendung. Die Wörter *null* bis *neun* waren je einmal von 100 Sprecherinnen und 100 Sprechern aufgesprochen worden. Trainings- und Testdatensatz waren disjunkt und umfassten jeweils 1000 Artikulationen und die Klassifizierung erfolgte *sprecherunabhängig*.

A. Anzahl und Art der notwendigen Merkmale

Es wurde untersucht, welche Anzahl von sekundären Merkmalen notwendig ist, um eine optimale Erkennungsleistung zu erzielen. In Abbildung 2 sind die Fehlerraten für unverrauschte Trainings- und Testdaten sowie für mit sprachsimulierenden Rauschen (CCITT G.227) bei einem Signal-Rausch-Abstand von 10dB gestörten Testdaten über der Anzahl der sekundären Merkmale aufgetragen.



a) saubere Trainings- und Testdaten (dunkel) *b) bei CCITT Rauschen mit 10dB SNR*

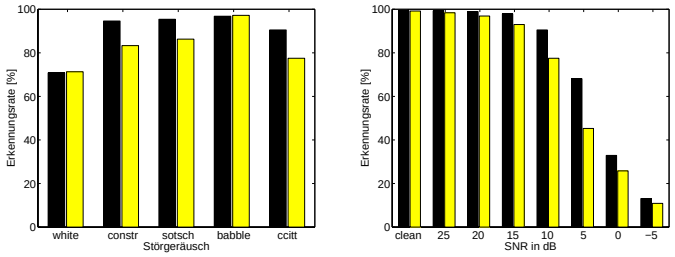
Abb. 2: *Fehlerraten in Abhängigkeit von der Anzahl der Merkmale*

Es ist zu erkennen, dass die Fehlerrate für saubere Testsignale erst bei unter 60 Merkmalen deutlich ansteigt. Der Fehler für leicht verrauschte Testdaten nimmt jedoch bereits bei unter 80 Merkmalen zu. Da die Größe der zu invertierenden Matrix $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ mit N^2 ansteigt, erscheint eine Anzahl von $N = 80$ Merkmalen ein sinnvoller Kompromiss zwischen Rechenaufwand und Effektivität zu sein. Für die folgenden Experimente wurde daher grundsätzlich mit 80 Merkmalen gearbeitet. Die freien Parameter eines neuen sekundären Merkmals wurden im obigen Experiment aus folgenden Intervallen zufällig gewählt: $f \in [1, 15]$, $f_0 \in [-14, 14]$ (alle Frequenzkanäle) und $t_0 \in [0, 30]$ (entspricht 300 ms) sowie $\Delta f \in [1, 5]$ und $\Delta t \in [1, 5]$. Damit ergeben sich ca. $2 \cdot 10^6$ mögliche Merkmale. Die Erkennungsleistung verändert sich nur geringfügig, wenn die Ausdehnung des großen Fensters auf ein Element der Matrix ($\Delta f = \Delta t = 1$) verringert wird. Wird weiterhin festgelegt, dass der Abstand der beiden Fenster in Zeit oder Frequenzbereich verschwinden muss, so ergeben sich eine Vervierfachung der Fehlerrate auf ungestörten Testdaten im ersten und eine Verdoppelung der Fehlerrate auf verrauschten Testdaten (CCITT bei 10dB SNR) im zweiten Fall. Dies deutet darauf hin, dass bei gegebenem Aufbau für eine robuste Einzelworterkennung sekundäre Merkmale mit einer Zeit- *und* Frequenzübergreifenden Integration notwendig sind.

B. Robustheit

In einem weiteren Experiment wurde die Robustheit des PEMO/FFNN Systems weiter untersucht und die Ergebnisse mit dem Erkennungssystem eines Referenzerkenners aus PEMO und lokal-rekurrentem Neuronalen Netz (LRNN) nach Kasper et al. [kas97] verglichen. Die Kombination PEMO/LRNN hat sich bereits als besonders robustes Einzelworterkennungssystem herausgestellt [tch97]. Addiert wurden bei definierten SNR neben CCITT-Rauschen außerdem weißes Rauschen (WHITE), weiteres sprachsimulierendes Rauschen[1] (SOTSCH) sowie Baustellenlärm[2] (CONSTR) und überlagerte Sprachgeräusche mehrerer Sprecher[3] (BABBLE).

[1] Kollmeier, B. et al., Audiol. Akustik 27, 1988
[2] Siemens 1992: *CDROM for fitting of hearing programs*
[3] NOISEX Datenbank, Varga, A. et al., tech. report 1992



a) additive Störgeräusche bei 10dB SNR *b) CCITT Rauschen bei veschiedenen SNR*

Abb. 3: *Erkennungsraten für FFNN (dunkel) und LRNN (hell)*

Abbildung 3 a) zeigt die Erkennungsraten für PEMO/LRNN und PEMO/FFNN bei verschiedenen additiven Störgeräuschen jeweils mit einem SNR von 10dB. In allen Fällen erreicht das neue System aus PEMO/FFNN eine vergleichbare oder bessere Erkennungsrate als das bisher robusteste Referenzsystem. In Abbildung 3 b) sind die Erkennungsraten für PEMO/LRNN und PEMO/FFNN bei additivem CCITT Rauschen und verschiedenen Signal-Rausch-Verhältnissen (SNR) aufgetragen. Es wird deutlich, dass die Verwendung des PEMO/FFNN Erkenners gegenüber dem schon robusten Referenzsystem PEMO/LRNN noch eine signifikante Erhöhung der Erkennungsrate bewirkt.

V. Zusammenfassung und Ausblick

Es konnte gezeigt werden, dass die vorgestellte Kombination aus perzeptiver Vorverarbeitung mit dem Perzeptionsmodell nach Dau et al. [dau96] und dem *Feature Finding Neural Network* nach Gramß eine gegenüber additiven Störgeräuschen robuste Klassifikation von isolierten Einzelwörtern ermöglicht. Dabei erlaubt das System aus PEMO/FFNN eine weitere Verbesserung der Robustheit gegenüber bekannten robusten Systemen mit rekurrenten neuronalen Netzen. Bei einer Reduzierung der erlaubten sekundären Merkmale auf reine Frequenz- oder Zeitverarbeitung sinkt die Klassifikationsleistung. Dies deutet auf die Notwendigkeit von spektral und temporal ausgedehnten sekundären Merkmalen hin, ist jedoch ebenso wie die genaue Beschaffenheit der optimalen Merkmale Gegenstand weiterer Experimente. Für die beschriebenen Spracherkennungsexperimente mit bereits segmentierten Einzelwörtern konnten die sekundären Merkmale über den gesamten Zeitverlauf integriert werden. Zur Erkennung kontinuierlicher Sprache kann und muss darauf verzichtet werden. Auch dies ist Objekt weiterer Forschung.

Bedanken möchten wir uns bei Klaus Kasper und Herbert Reininger von der Universität Frankfurt dafür, dass sie uns ihre LRNN Implementation zur Benutzung überlassen haben.

Literatur

[all94] J.B. Allan: *How Do Humans Process and Recognize Speech*, IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 567-576, 1994.

[dau96] T. Dau, D. Püschel, A. Kohlrausch: *A quantitive model of the "effective" signal processing in the auditory system I+II.*, JASA 99 (6), pp. 3615-3631, 1996.

[gra92] T. Gramß: *Worterkennung mit einem künstlichen neuronalen Netzwerk*, Dissertation, Universität Göttingen, 1992.

[kae00] C. Kaernbach : *Early auditory feature coding* Contr. 8th Oldenburg Symposium on Psychol. Acoustics. Edited by August Schick u.a., BIS, Universität Oldenburg. pp. 295–308, 2000.

[kas97] K.Kasper, H. Reininger, D. Wolf: *Exploiting the Potential of Auditory Preprocessing for Robust Speech Recognition by LRNN*, Proc. ICASSP vol. 2, pp. 1223–1227, 1997.

[kle98] M. Kleinschmidt, J. Tchorz, T. Wittkop, V. Hohmann, B. Kollmeier: *Robuste Spracherkennung durch binaurale Richtungsfilterung und gehörgerechte Vorverarbeitung*, Fortschritte der Akustik - DAGA, pp. 396–397, 1998.

[tch97] J.Tchorz, K.Kasper, H. Reininger, B. Kollmeier: *On the Interplay between Auditory-based Features and Locally Recurrent Neural Networks for Robust Speech Recognition in Noise*, Eurospeech 97, pp. 2075–2078, ESCA, Patras, Greece, 1997.

[tch99] J. Tchorz, B. Kollmeier: *A Model of Auditory Perception as Front End for Automatic Speech Recognition*, J. Acoust. Soc. Am., vol. 106, no. 4, pp. 2040–2050, 1999.