Spracherkennung und Sprechererkennung mit perzeptionsorientierter Merkmalsextraktion

Herbert Reininger

Institut für Angewandte Physik der J.W. Goethe-Universität Frankfurt a. M.
Robert-Mayer-Straße 2-4, D-60054 Frankfurt am Main, BRD
e-mail: herbrein@apx00.physik.uni-frankfurt.de

1 Einleitung

Ein grundlegendes Problem automatischer Sprach- und Sprechererkennungssysteme ist die mangelnde Robustheit der Erkennungsleistung gegenüber Unterschieden zwischen den Aufnahmebedingungen der Trainingsdaten und denen der zu erkennenden Sprachsignale. Ein Lösungsansatz besteht darin, mittels perzeptionsorientierten Verfahren robuste Merkmalsvektoren zu finden, deren charakteristische Ausprägung und Dynamik weniger sensitiv gegen veränderte akustische Bedingungen sind.

Im Beitrag werden verschiedene derartige Merkmalsextraktionsverfahren (MFCC, PLP, RASTA, PEMO) diskutiert und die Leistungsfhigkeit dieser Verfahren im Kontext der sprecherunabhängigen Worterkennung von geräuschbehafteten Sprachsignalen verglichen.

2 Verfahren der perzeptionsorientierten Merkmalsextraktion

2.1 Mel Frequency Cepstral Coefficients (MFCC)

Die Parametrisierung von Sprache mit MFCC besteht in der Simulation einer mittels FFT realisierten, auditorischen Filterbank, deren Filter äquidistant entlang der Mel-Skala angeordnet sind. Zur Gewichtung der Spektralkoeffizienten werden Dreiecksfunktionen verwendet, die für eine Mittenfrequenz den Gewichtswert $g_i = 1.0$ und für die benachbarten Mittenfrequenzen einen Gewichtswert $g_i = 0$ besitzen. Die akkumulierten Leistungen der einzelnen Frequenzbänder werden logarithmiert und danach mit Hilfe der Cosinus-Transformation in Cepstralkoeffizienten umgerechnet.

2.2 Perceptual Linear Prediction (PLP)-basierte Verfahren

PLP-Verfahren unterscheiden sich von MFCC durch eine komplexere Gestaltung der spektralen Gewichtungsfunktion und durch eine autoregressive Modellierung [1]. Zur Berechnung der Leistung der einzelnen Frequenzkanäle werden trapezförmige Gewichtungsfunktionen verwendet und anschließend die Leistung jedes Frequenzkanals entsprechend einer Kurve gleicher Lautheit bei 40 dB gewichtet. Die Kompression der Leistung erfolgt gemäß $y(t) = \sqrt[3]{x(t)}$. Ausgehend von der gewichteten und komprimierten Leistung in den Frequenzbändern werden zunächst Autokorrelations-Koeffizienten berechnet, die dann mittels LPC-Methoden in Cepstralkoeffizienten transformiert werden. Typischerweise werden wesentlich weniger Autokorrelations-Koeffizienten (i.allg. 9) berechnet als kritische Frequenzbänder vorliegen.

Zur Steigerung der Robustheit von PLP-Koeffizienten gegenüber Störungen wurde die Begrenzung der Modulationsfrequenzen der einzelnen Frequenzkanäle eingesetzt. Bei dem als RASTA (*RelAtive SpecTrA*) [2] bezeichneten Verfahren werden mit einem Bandpaßfilter die Modulationsfrequenzen der einzelnen Kanäle auf den Bereich von ca. 1 Hz bis ca. 10 Hz beschränkt. Nicht-sprachliche Anteile aus einem Sprachsignal werden hierdurch unterdrückt.

Vor der Filterung der Spektralkoeffizienten werden diese logarithmisch komprimiert und nach der Filterung wieder expandiert. Danach erfolgt eine Verarbeitung analog der zur Berechnung von PLP-Koeffizienten. Dieses Verfahren wird als LOG-RASTA bezeichnet. Da spektrale Verzerrungen des Sprachsignals als konstante additive Komponenten im logarithmierten Leistungsspektrum erscheinen, werden derartige Störungen durch die RASTA-Filterung eliminiert. Zudem zeigte sich, daß mit RASTA auch psychoakustische Befunde, wie die zeitliche Verdeckung, modelliert werden [3].

Um die geringe Robustheit von LOG-RASTA gegenüber additiven Störgeräuschen zu verbessern, wurde anstelle einer logarithmischen Abbildung eine adaptive Nichtlinearität, gesteuert durch einen geräuschabhängigen Faktor J eingeführt, was als JAH-RASTA bezeichnet wird.

2.3 Perzeptionsmodell PEMO

Das Perzeptionsmodell PEMO wurde ursprünglich zur Beschreibung psychoakustischer Ergebnisse, wie Verdeckung im Zeit- und Frequenzbereich, in Oldenburg und Göttingen entwickelt [4]. Es handelt sich bei PEMO um ein Verfahren, das im Zeitbereich arbeitet. Nach Präemphase erfolgt eine Frequenzzerlegung mit Hilfe einer Gammaton-Filterbank, wobei die Bandpässe der Filterbank äquidistant auf der ERB-Skala angeordnet sind. Zur Zerlegung von Telefonsprache werden 17 Frequenzbänder verwendet. Die Teilbandsignale werden gleichgerichtet und einer Tiefpaß-Filterung mit Eckfrequenz von 1 kHz unterzogen, womit Eigenschaften der Haarzellen simuliert werden. Die Kompression der Audiosignale durch das Gehör wird modelliert durch 5 aufeinanderfolgende Adaptionsschleifen, die jeweils aus einem RC-Tiefpaß und einem Dividierer bestehen. Die Tiefpaßfilter haben Relaxationszeiten zwischen 5 ms und 500 ms. Im stationären Zustand wird durch die Adaptionsschleifen ein Teilbandsignal $x(n)$ gemäß $\sqrt[3.2]{x(n)}$ komprimiert. Schnelle Änderungen des Eingangssignals werden hingegen verstärkt. Auf diese Weise wird eine starke Kontrastierung von Signaländerungen erreicht. Diese Art der adaptiven Kompression stellt eine signifikante Differenz zu den bisher dargestellten Verfahren zur Merkmalsextraktion dar. Schlielich wird ein Tiefpaßfilter mit einer Eckfrequenz von 8 Hz auf jedes Ausgangssignal einer Adaptionsschleife angewandt. Die Koeffizienten der Merkmalsvektoren entstehen durch Mittelung ber eine Dauer von 10 ms.

3 Experimente zur robusten Spracherkennung

Die Experimente wurden am ZifKom-Korpus durchgeführt, wobei die Äußerungen der 10 Zahlwörter *Null*, *Eins*, *Zwo*, *Drei*, *Vier*, *Fünf*, *Sechs*, *Sieben*, *Acht* und *Neun* von 200 Sprechern verwendet wurden. Die Äußerungen von 100 Sprechern wurden zum Training der Spracherkennungssysteme (SES) verwendet und die Äußerungen der restlichen Sprecher zum Test. Das Training der SES erfolgte mit ungestörten Sprachsignalen, deren Frequenzbereich auf Telefonbandbreite begrenzt war. Zur Messung der Leistungsfähigkeit der SES wurden sowohl die ungestörten Testdaten (cl) als auch geräuschbehaftete Sprachdaten verwendet. Zur Simulation von typischen Störsituatio-

nen wurden die Daten der Testmenge mit weißem Rauschen (WR), sprachsimulierendem Rauschen (SR) oder einem typischen Baustellengeräusch (BG) additiv überlagert. Mit Hilfe der Störgeräusche wurden jeweils Testmengen mit einem Signal-Rausch-Abstand (SNR) von 0 dB, 10 dB und 20 dB hergestellt.

Eine Merkmalsextraktion erfolgte im Abstand von 10 ms. Für alle Parametrisierungen mit Ausnahme von PEMO wurde ein Koeffizient zur Charakterisierung der Kurzzeitenergie verwendet. Damit ergaben sich für LPC-CEP und MFCC jeweils 13 Basiskoeffizienten, für PLP, LOG-RASTA und JAH-RASTA jeweils 9 und für PEMO [1] 17 Basiskoeffizienten.

Die Robustheit eines SES ist das Resultat eines komplexen Zusammenwirkens der Verfahren zur Merkmalsextraktion und der zur Realisierung des Bewertungsmoduls angewandten Modellierungstechnik. Vor diesem Hintergrund wurde jeweils ein auf kontinuierlichen Hidden-Markov-Modellen (CHMM)-basiertes und ein auf Lokal Rekurrenten Neuronalen Netzen (LRNN) basiertes SES für jede Merkmalsart realisiert. Die CHMM-basierten SES wurden aus Wortmodellen aufgebaut, die jeweils 8 emittierende Zustände mit je 5 Gaußdichten haben. Die diagonale Kovarianzmatrix wurde für alle Dichten identisch gewählt und anhand der Trainingsdaten berechnet. Für alle CHMM-basierten SES wurden neben den Basiskoeffizienten auch Deltakoeffizienten verwendet. Die LRNN-basierten SES wurden mit Supervektoren der Länge $L_S = 5$ betrieben. Zum Aufbau der Supervektoren wurden ausschließlich Basiskoeffizienten verwendet. In der versteckten Schicht wurden 13 x 13 Neurone angeordnet, die über eine Nachbarschaft von $n = 5$ verfügten. Zur Repräsentation der Wörter des Vokabulars wurden 10 Ausgangsneurone verwendet.

Tabelle 1: Prozentuale Worterkennungsraten

SES	cl	WR		SR		BG	
SNR	/	0	10	0	10	0	10
LPC-CEP							
CHMM	99.3	12.9	33.0	24.5	83.4	30.1	82.6
LRNN	98.9	17.5	32.8	15.5	22.4	13.5	28.1
MFCC							
CHMM	98.8	18.8	63.5	12.8	86.0	18.1	83.7
LRNN	97.2	10.7	24.7	10.0	23.9	11.2	25.8
PLP							
CHMM	99.2	9.5	25.2	11.6	56.0	13.3	70.0
LRNN	98.4	11.8	31.0	10.0	21.5	10.0	21.0
LOG-RASTA							
CHMM	98.9	21.0	73.3	17.0	71.5	24.1	75.8
LRNN	99.5	6.5	43.7	10.2	36.8	6.7	44.0
JAH-RASTA							
CHMM	99.2	55.9	82.8	42.5	82.3	51.4	84.1
LRNN	97.3	41.3	66.7	29.2	64.6	24.0	66.6
PEMO							
CHMM	96.4	24.4	71.8	43.0	89.0	26.0	79.9
LRNN	97.6	51.1	90.7	49.7	96.5	60.3	94.7
PEMO-CEP							
CHMM	97.6	55.5	85.7	74.8	93.3	67.3	91.6
LRNN	98.6	51.8	90.1	44.9	94.2	56.2	95.1

4 Diskussion der Ergebnisse

In Tabelle 1 sind die gemessenen prozentualen Worterkennungsraten zusammengestellt. Eine Analyse der Meßergebnisse läßt erkennen, daß im Kontext der ausgeführten Experimente LPC-CEP, MFCC und PLP eine ähnlich geringe Robustheit aufweisen. Eine Verbesserung der Robustheit zeigt sich bei Verwendung von LOG-RASTA zur Parametrisierung der Sprache. Zu einer deutlich erhöhten Robustheit führt die Kombination von JAH-RASTA mit LRNN oder CHMM, wobei bei einer Kombination mit CHMM wesentlich höhere absolute Raten erzielt werden konnten. Hierbei ist jedoch zu bedenken, daß zur Adaption des JAH-Wertes eine Schätzung der Störleistung durchgeführt werden muß und hierfür die ersten 100 ms der zu erkennenden Äußerung keine Sprache enthalten dürfen. Ob dies allgemein gewährleistet werden kann ist fraglich.

Entgegen der Tendenz, daß LRNN-basierte SES sehr sensitiv gegenüber gestörten Sprachsignalen sind, konnte mit der Kombination von PEMO mit LRNN eine extrem gute Robustheit insbesondere gegenüber additivem Störgeräusch erreicht werden [5,6]. Im Vergleich dazu ist die Leistungsfähigkeit der Kombination von PEMO mit CHMM gering. Selbst an ungestörten Daten werden nur ca. 96% Erkennungsrate erreicht. Eine Analyse dieses Befundes zeigt, daß die Kontrastierung der Merkmale bei PEMO zu Verteilungen führt, die sich nicht für eine Modellierung mit HMM eignen. Durch Transformation der PEMO-Merkmale in den Cepstralbereich (PEMO-CEP) gelingt es, eine für HMM-Modellierung adäquate Darstellung zu finden. Mit PEMO-CEP erreichen CHMM die gleich hohe Robustheit, wie mit JAH-RASTA. LRNN erzielen die mit PEMO gewonnene Robustheit auch mit PEMO-CEP, obwohl die Anzahl der Merkmalskomponenten nahezu halbiert ist.

[1] Die Durchführung der Experimente erfolgte mit Hilfe einer von der AG Medizinische Physik der Universität Oldenburg zur Verfügung gestellten Implementierung.

5 Zusammenfassung

Die Integration von einfachen psychoakustischen Befunden, wie die Zusammenfassung von Frequenzen zu Frequenzbändern oder die Berücksichtigung von Kurven gleicher Lautheit, läßt keine Steigerung der Robustheit von SES erwarten. Eine signifikante Verbesserung der Robustheit von SES wird erst durch eine konstrastierende Filterung erreicht, wie sie in den RASTA-Verfahren und in PEMO realisiert ist.

Literatur

[1] Hermansky, H.: *Perceptual Linear Predictive (PLP) Analysis of Speech*, J. Acoust. Soc. Am., vol 87, 1990, S. 1738-1752

[2] Hermansky, H., Morgan, N.: *Rasta Processing of Speech*, IEEE Trans. on Speech and Audio Processing, vol. 2, 1994, S. 578-589

[3] Hermansky, H.: *Should recognizers have ears*, Speech Communication, vol 25, 1998, S. 3-27

[4] Dau, T., Püschel, D., Kohlrausch, A.: *A quantitave model of the 'effective' signal processing in the auditory system*, Part I and II in J. Acoust. Soc. Am., vol 99, 1996, S. 3615-3631

[5] Tschorz, J., Kleinschmidt, M., Kasper, K., Kollmeier, B.: *Auditory feature extraction and recognizer dependencies*, Proc. of Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere 1999, im Druck

[6] Kasper, K., Reininger, H.: *Evaluation of PEMO in Robust Speech Recognition*, Proc. of ASA, Berlin 1999