Diskriminanzanalyse zur differenzierenden Erkennung leicht verwechselbarer Klassen

Hans-Günter Hirsch, Fabian Schmitt

Institut für Mustererkennung, Hochschule Niederrhein, 47805 Krefeld,
E-Mail:hans-guenter.hirsch@hs-niederrhein.de

Einleitung

Im Fall der automatischen Erkennung gestörter Sprachsignale beobachtet man bei der Festlegung des Inhalts eines sprachlichen Abschnitts häufig die Bestimmung nahezu gleich großer Wahrscheinlichkeiten für mehrere Klassen des zur Erkennung verwendeten Vokabulars. Im Rahmen der hier vorgestellten Untersuchungen wird die Sprachdatenbasis TIDigits eingesetzt, die die Aufnahmen englischer Ziffernketten beinhaltet. Die Ziffern werden mit Hidden Markov Modellen (HMM) beschrieben, so dass ein sprachlicher Abschnitt einer Ziffer entspricht. Die Entscheidung an Hand der maximalen Wahrscheinlichkeit führt bei mehreren Klassen, für die eine nahezu gleich große Wahrscheinlichkeit berechnet wird, zu einem eher zufällig richtigem oder falschem Ergebnis. Der im Folgenden vorgestellte Ansatz versucht in diesen kritischen Fällen durch eine weitergehende differenzierende Analyse und Erkennung die Entscheidungsgrundlage zu verbessern.

Differenzierende Analyse und Erkennung

Zunächst wird zur Extraktion akustischer Merkmale eine robuste MEL Cepstralanalyse eingesetzt. Dabei werden mit Hilfe einer Diskreten Cosinus Transformation (DCT) des MEL Spektrums die zugehörigen Cepstralkoeffizienten bestimmt. Das MEL Spektrum besteht dabei aus 24 Werten, aus denen die 12 Cepstralkoeffizienten C_1 bis C_{12} berechnet werden. Der hier vorgestellte Ansatz beruht auf einem Ersatz der DCT durch eine Hauptkomponentenanalyse (PCA) oder eine lineare Diskriminanzanalyse (LDA). Gerade die LDA verfolgt das Ziel Merkmalswerte zu generieren, um zwei oder eventuell auch mehrere Klassen besser differenzieren zu können. Der gesamte Erkennungsvorgang besteht aus einer ersten Erkennungsstufe unter Verwendung der üblichen MEL Cepstralkoeffizienten. Werden dabei Abschnitte detektiert, für die in etwa gleich große Wahrscheinlichkeiten für mehrere Ziffernmodelle bestimmt werden, so erfolgt eine zweite differenzierende Erkennung in einer zweiten Analyse- und Erkennungsstufe. Dabei werden die logarithmierten MEL Spektren mit einer speziell zur Separierung zweier Klassen konzipierten LDA oder PCA transformiert. Die PCA oder LDA Matrix wird dabei aus den ungestörten TIDigits Trainingsdaten mit einer Beschränkung auf die MEL Spektren der beiden zu vergleichenden Klassen bestimmt. Es werden separate Transformationsmatrizen bestimmt, um jeweils zwei Klassen (Ziffern) besser differenzieren zu können. Um die Differenzierung der beiden Klassen noch weiter zu verbessern, werden mehrere Transformationsmatrizen in Abhängigkeit der zeitlichen Lage innerhalb einer Ziffer bestimmt. Zur Bestimmung dieser Matrizen werden in der Trainingsphase alle MEL Spektren einer Klasse mit Hilfe eines Clusteralgorithmus in eine vorgegebene Anzahl von Segmenten unterteilt, wobei sich für die im Rahmen dieser Untersuchungen verwendeten sprachlichen Abschnitte eine Wahl von 3 oder 4 Segmenten als geeignet zu erweisen scheint. Aus allen Spektren, die einem zeitlichen Segment zugeordnet werden, werden die Parameter eines Gaussian Mixture Modells (GMM) bestimmt. Mit Hilfe dieser GMMs kann dann in der Erkennungsphase die Zuordnung jedes Spektrums einer zu erkennenden Folge von MEL Spektren zu den einzelnen Segmenten vorgenommen werden. Dies wird beispielhaft in Abbildung 1 für die beiden Klassen der Ziffern „FIVE“ und „NINE“, für die häufiger eine etwa gleich große Wahrscheinlichkeit berechnet wird, veranschaulicht.

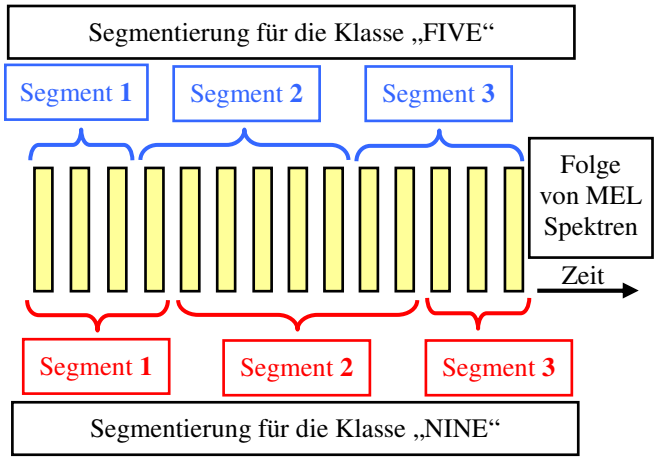

Abbildung 1: Klassenspezifische Segmentierung einer Folge von MEL Spektren.

In der Mitte der Abbildung wird die Folge der MEL Spektren, die aus der Analyse der Äußerung einer Ziffer bestimmt werden, veranschaulicht. Darüber und darunter ist die Unterteilung der Folge in 3 Segmente dargestellt, die individuell mit den GMMs einer jeden Klasse mittels der Bestimmung des GMMs mit maximaler Wahrscheinlichkeit vorgenommen wird. In der Trainingsphase wird für jede mögliche Segmentkombination, in dem dargestellten Beispiel mit 3 Segmenten sind es neun Kombinationen, eine individuelle LDA bzw. PCA Matrix unter Verwendung der MEL Spektren der beiden Klassen, die dem jeweiligen Segment zugeordnet wurden, durchgeführt. Damit werden insgesamt zum Beispiel bei Betrachtung von 10 Klassen für die Ziffern und eine Unterteilung in 3 Segmente 405 LDA bzw. PCA Matrizen benötigt. Die Zahl von 405 Matrizen resultiert aus 45 Klassenvergleichen, multipliziert mit den 9 Segmentkombinationen. Die Zahl von 45 beschreibt die Anzahl von möglichen Vergleichen unterschiedlicher Klassen bei insgesamt 10 Klassen.

In der in Abbildung 1 dargestellten Folge von Spektren wird beispielsweise für die Bestimmung der differenzierenden Merkmale des ersten Vektors die Matrix zum Vergleich des jeweils ersten Segments der beiden Klassen „FIVE“ und „NINE“ benötigt. Aus der klassenspezifischen Zuordnung jedes weiteren MEL Spektrums zu einem der 3 Segmente ergeben sich für dieses Spektrum die entsprechenden Indizes zur Auswahl der Transformationsmatrix. Um mit den aus der Transformation resultierenden Merkmalen die Berechnung einer Wahrscheinlichkeit vorzunehmen, werden in der Trainingsphase aus allen differenzierenden Merkmalsvektoren, die für jedes individuelle Paar von zu vergleichenden Klassen und jede Segmentkombination bestimmt wurden, die Parameter eines GMMs für jede Klasse berechnet. Für die beispielhafte Betrachtung von 10 Klassen und die Unterteilung in 3 Segmente ergeben sich insgesamt 810 GMMs, resultierend aus der Multiplikation von 90 Klassenkombinationen und den 9 Segmentkombinationen. Die Zahl von 90 ergibt sich aus dem Vergleich jeder Klasse zu allen anderen außer sich selbst. Mit diesen GMMs kann in der Erkennungsphase durch die multiplikative Verknüpfung der für jeden Vektor einer Folge berechneten Wahrscheinlichkeiten die Gesamtwahrscheinlichkeit für die Zuordnung zu einer Klasse berechnet werden.

Mit der zuvor beschriebenen Vorgehensweise wird in der zweiten Erkennungsstufe ein Vergleich jeder Klasse, für die eine hohe Wahrscheinlichkeit in der ersten Erkennungsstufe berechnet wurde, zu allen anderen Klassen mit hoher Wahrscheinlichkeit durchgeführt. Als Ergebnis erhält man eine Matrix von Wahrscheinlichkeiten, die man wiederum zur Bestimmung eines differenzierenden Erkennungsergebnisses für einen in der ersten Stufe unzuverlässig erkannten Abschnitt verwenden kann. In Tabelle 1 wird diese Matrix beispielhaft veranschaulicht, die sich bei Bestimmung der Klassen „5“, „9“, „1“ und „8“ als nahezu gleich wahrscheinlich in der ersten Stufe ergibt. An den mit x markierten Stellen tritt jeweils ein für diesen Klassenvergleich individuell berechneter Wahrscheinlichkeitswert auf.

Tabelle 1: Wahrscheinlichkeitsmatrix als Ergebnis der differenzierenden Erkennung der mit hoher Wahrscheinlichkeit erkannten Klassen

		Vergleichsklasse			
		5	9	1	8
als erkannt angenommene Klasse	5	–	x	x	x
	9	x	–	x	x
	1	x	x	–	x
	8	x	x	x	–

Um zu einem Gesamterkennungsergebnis zu kommen, kann man die mittlere Wahrscheinlichkeit des Vergleichs einer jeden Klasse zu allen anderen Klassen berechnen. Dies entspricht einer zeilenweisen Mittelwertbildung der in Tabelle 1 angedeuteten Wahrscheinlichkeitswerte. Mit Hilfe der größten mittleren Wahrscheinlichkeit erhält man dann ein finales Erkennungsergebnis. Daneben sind auch alternative Möglichkeiten der Auswertung aller Werte in der Matrix und damit der Bestimmung eines Endergebnisses denkbar, die im späteren Verlauf dieses Projekts untersucht werden sollen.

Anwendung der differenzierenden Erkennung

Die Idee der differenzierenden Analyse und Erkennung wird im Rahmen eines Projekts zur Verbesserung der Erkennung gestörter Sprachsignale eingesetzt. Im Rahmen dieses Projekts wird eine alternative Vorgehensweise zur Berechnung der Wahrscheinlichkeit, die die Erzeugung einer beobachteten Folge von Merkmalsvektoren mit einem HMM oder einer Folge von HMMs beschreibt, untersucht. Üblicherweise wird diese Wahrscheinlichkeit gemäß dem zeitlich sequentiellen Auftreten der in der Sprachanalyse bestimmten Merkmalsvektoren berechnet. Bei dem hier untersuchten Ansatz erfolgt die Berechnung ausgehend von den zeitlichen Abschnitten, in denen das Signal-zu-Rauschleistungsverhältnis (SNR) hohe Werte annimmt. Ausgehend von dem Zentrum der detektierten Abschnitte erfolgt eine separate Wahrscheinlichkeitsberechnung, zeitlich vorwärts und zeitlich rückwärts gerichtet, wie es in Abbildung 2 veranschaulicht wird.

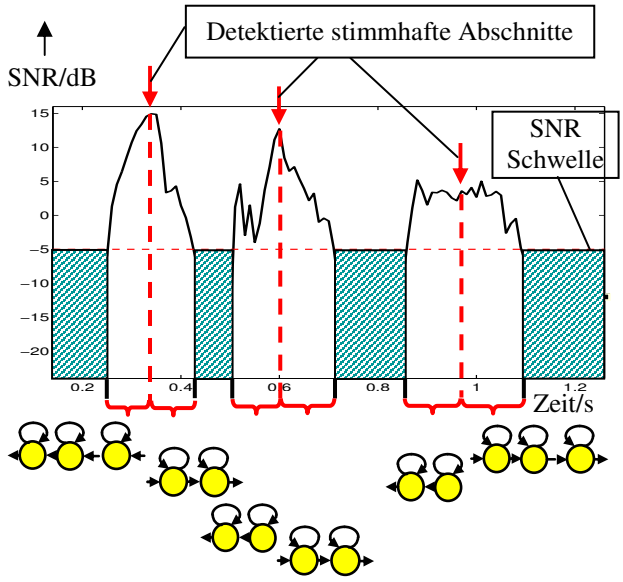

Abbildung 2: Ergebnis der Detektion stimmhafter Abschnitte für eine aus 3 Ziffern bestehende Äußerung und Modellierung der Äußerung mit zeitlich rückwärts und vorwärts gerichteten HMMs.

In Abbildung 2 ist der Verlauf des SNRs eines gestörten Signals, das drei Ziffern beinhaltet, dargestellt. Das Ergebnis der Detektion der Bereiche mit hohem SNR wird visualisiert. In den Abschnitten mit hohem SNR treten in der Regel stimmhafte Laute mit hoher Energie auf. Daher wird zunächst ein Verfahren zur Detektion der stimmhaften Bereiche eingesetzt [1]. Ausgehend von dem Merkmalsvektor im Zentrum eines detektierten stimmhaften Abschnitts werden die beiden Wahrscheinlichkeiten berechnet, die vor dem stimmhaften Zentrum auftretenden Merkmalsvektoren mit einem speziellen zugehörigen HMM bzw. die danach auftretenden Vektoren mit einem weiteren

HMM zu erzeugen. Dazu werden in der Trainingsphase für die ungestörten Signale der TIDigits Trainingsdaten entsprechende Teilwort HMMs bestimmt, um das zeitlich rückwärts gerichtete Auftreten der Merkmalsvektoren vor einem stimmhaften Laut bzw. das zeitlich vorwärts gerichtete Auftreten der Merkmalsvektoren nach dem stimmhaften Laut zu modellieren. Des Weiteren kann die Betrachtung auf die Merkmalsvektoren beschränkt werden, die ein SNR oberhalb einer bestimmten Schwelle besitzen, wie es in Abbildung 2 dargestellt ist. Damit ergibt sich für die drei detektierten Bereiche jeweils eine zeitliche Eingrenzung.

Die in dem Projekt mit dieser Vorgehensweise bisher erzielten Erkennungsergebnisse [2] deuten an, dass bei einer alleinigen Betrachtung dieses Ansatzes keine wesentlichen Verbesserungen gegenüber einer herkömmlichen Erkennung erzielt werden können. Eine Schwierigkeit besteht beispielsweise in der korrekten Ermittlung aller stimmhaften Abschnitte bei gestörten Signalen. Allerdings zeigt sich ein Potential durch die Verknüpfung einer herkömmlichen Erkennung mit diesem neuen Ansatz die Erkennung gestörter Sprachsignale zu verbessern. Aus der Analyse der bisher erzielten Ergebnisse entstand die Idee der differenzierenden Analyse und Erkennung, wie sie im vorherigen Abschnitt beschrieben wurde.

Der erste, im Rahmen des Projekts untersuchte Ansatz basiert auf einer herkömmlichen Erkennung, bei der die robusten MEL Cepstralkoeffizienten in Kombination mit Teilwortmodellen, mit denen die Ziffernteile vor und nach einem stimmhaften Laut modelliert werden, verwendet werden. Im Gegensatz zur zuvor beschriebenen Vorgehensweise in der Trainingsphase wird das Modell für den vorderen Ziffernteil allerdings aus den zeitlich vorwärts gerichteten Folgen von Merkmalsvektoren bestimmt, um die herkömmliche, zeitlich vorwärts gerichtete Erkennung zu ermöglichen. Dabei stellen sich für verschiedene Aufnahmebedingungen die in den Tabellen 2 und 3 aufgeführten Wortfehlerraten ein.

Tabelle 2: Wortfehlerraten im Innern eines Kraftfahrzeugs

clean	car15dB	car10dB	car5dB	car0dB
0,5 %	1,3 %	2,4 %	6,5 %	18,7 %

Tabelle 3: Wortfehlerraten in räumlichen Umgebungen

clean	int15dB	int10dB	int5dB
0,5 %	3,1 %	7,1 %	17,0 %

Mit der Aufnahmebedingung „clean“ werden die ungestörten TIDigits des zum Test von Erkennungssystemen vorgesehenen Teils der Datenbasis referenziert, der aus 8700 Aufnahmen mit insgesamt etwa 28000 Ziffern besteht. Die mit den Termen „car“ bzw. „int“ versehenen Aufnahmesituationen beinhalten die gleichen Testdaten, denen allerdings entweder ein Störgeräusch im Innern eines Kraftfahrzeugs (car) bzw. ein Störgeräusch aus verschiedenen räumlichen Situationen (int als Kürzel für interior), z.B. in einem Einkaufszentrum oder in einem Restaurant, additiv überlagert werden [3]. Der Zahlenwert hinter dem jeweiligen Umgebungsterm definiert das SNR in dB und damit den Faktor, mit dem das Störgeräusch gewichtet zu dem Sprachsignal addiert wird. Die gestörten Aufnahmen entstammen der Datenbasis mit der Bezeichnung „Aurora5“ [4]. Man beobachtet zum einen die bekannte Erhöhung der Fehlerrate bei kleiner werdendem SNR sowie die prinzipiell höheren Fehlerraten in räumlichen Störsituationen. Das eher als stationär anzusehende Störgeräusch im Innern eines Autos kann durch die in dem Merkmalsextraktionsverfahren enthaltene adaptive Filterung besser kompensiert werden als die Störgeräusche in räumlichen Umgebungen, die häufig nicht stationäre Signalanteile beinhalten. Die sich unter Verwendung der Teilwortmodelle einstellenden Wortfehlerraten sind nahezu identisch mit den Fehlerraten, die bei Einsatz der üblichen Ganzwortmodelle erzielt werden. Dies zeigt, dass die Verwendung von Modellen, mit denen die Wortteile vor und nach einem stimmhaften Laut separat modelliert werden, zu keiner prinzipiellen Verschlechterung führt.

Die Idee der differenzierenden Analyse und Erkennung wird dann in einem zweiten Schritt jeweils separat auf die MEL Spektren des vorderen und des hinteren Wortteils angewendet. Zur zeitlichen Zuordnung der MEL Spektren zu den einzelnen Ziffernteilen wird das Ergebnis der herkömmlichen Erkennung herangezogen. Im Gegensatz zu der im vorherigen Abschnitt beschriebenen Vorgehensweise werden als sprachliche Einheit nicht ganze Ziffern, sondern die sich bei Aufspaltung im stimmhaften Bereich ergebenden Wortteile verwendet. Dabei treten bis auf die kürzere mittlere zeitliche Länge der Ziffernteile keine prinzipiellen Unterschiede zur Betrachtung ganzer Ziffern auf.

In einem ersten Erkennungsexperiment wird die eigentlich angedachte Auswertung der in der ersten Erkennungsstufe berechneten Wahrscheinlichkeiten und die Beschränkung auf die Klassen, für die die höchsten Wahrscheinlichkeiten berechnet wurden, außer Acht gelassen. Es soll das Potential bestimmt werden, alleine mit der differenzierenden Analyse und Erkennung eine Erkennung der Ziffernketten durchzuführen. Es wird eine differenzierende Erkennung für einen Vergleich jeder Klassen zu allen anderen vorgenommen. Damit erhält man eine Matrix von Wahrscheinlichkeitswerten, wie sie in Tabelle 1 dargestellt wurde, allerdings unter Berücksichtigung aller zur Erkennung verwendeten Klassen. Wie zuvor beschrieben, erfolgt die Zuordnung eines Ziffernteils zu der Klasse mit der höchsten mittleren Wahrscheinlichkeit in einer Zeile der Matrix. Für die Erkennung der ungestörten TIDigits Testdaten, mit denen die ersten Experimente durchgeführt wurden, stellen sich dabei die in Tabelle 4 aufgeführten Wortfehlerraten ein. Zum einen wird bei den Experimenten die Unterteilung in zwei bis vier Segmente variiert, zum anderen wird entweder die PCA oder die LDA verwendet. Bei Anwendung der LDA wird nur ein Koeffizient zur Differenzierung der beiden Klassen bestimmt. Die Transformationsmatrix besteht in diesem Fall nur aus einer Spalte.

Tabelle 4: Wortfehlerraten für verschiedene Parametrisierungen der differenzierenden Erkennung

Experiment	Fehlerrate
3 Segmente, 3 PCA Koeffizienten	5,07 %
3 Segmente, 10 PCA Koeffizienten	2,42 %
2 Segmente, 10 PCA Koeffizienten	3,92 %
3 Segmente, 15 PCA Koeffizienten	2,33 %
3 Segmente, 1 LDA Koeffizient	1,70 %
4 Segmente, 1 LDA Koeffizient	2,17 %

Bei der PCA wird die Anzahl der aus der Transformation resultierenden Werte im Bereich von 3 bis 15 variiert. Man beobachtet die geringste Fehlerrate bei Verwendung der LDA und einer Unterteilung in 3 Segmente. Es bestätigt sich damit die schon zu Beginn erwähnte Fähigkeit der LDA, im Vergleich zur PCA besser zur Differenzierung zweier Klassen geeignet zu sein.

Zusammenfassung und Ausblick

Es wird der Ansatz einer differenzierenden Analyse und Erkennung vorgestellt, der als zweite Stufe eines zweistufigen Erkennungssystems eingesetzt werden soll. Damit soll eine zuverlässigere Zuordnung von MEL Spektren zu einer von zwei oder mehreren Klassen, für die in der ersten Erkennungsstufe eine nahezu gleich große Wahrscheinlichkeit bestimmt wurde, erzielt werden. Im Rahmen der laufenden Untersuchungen wurden bisher zunächst Experimente angestellt, mit der differenzierenden Analyse und Erkennung als alleiniger Erkennungsstufe die ungestörten TIDigits zu erkennen. Für diesen Anwendungsfall, für den der Ansatz eigentlich nicht gedacht ist und auch nicht entwickelt wird, stellen sich schon recht hohe Worterkennungsraten ein, wobei die Fehlerrate höher ist im Vergleich zu einer herkömmlichen Erkennung unter Verwendung der Cepstralkoeffizienten.

Aber es deutet sich ein hohes Potential an, die differenzierende Aufgabe in einer zweiten Erkennungsstufe damit realisieren zu können. Dies soll in zukünftigen Experimenten untersucht werden. Zunächst wird dabei eine Detektion der als unzuverlässig einzustufenden Erkennungsergebnisse in der ersten Erkennungsstufe benötigt. Eine weitere Aufgabe besteht in der Bestimmung von LDA Matrizen, um die spektralen Merkmale im Fall gestörter Sprachsignale differenzieren zu können. Ein interessanter Ansatz zur Bestimmung von LDA Transformationen, mit denen die Erkennung gestörter Signale verbessert werden kann, findet sich in [5], der auch im Rahmen dieses Projekts angewendet und untersucht werden soll.

Danksagung

Die Autoren möchten sich bei der Deutschen Forschungsgemeinschaft (DFG) für die Unterstützung dieser Untersuchungen bedanken.

Literatur

[1] Hirsch, H.G., Kitzig, A., Kremer, F.: Detektion stimmhafter Abschnitte zur robusten Spracherkennung. Workshop Audiosignal- und Sprachverarbeitung im Rahmen der GI Jahrestagung Informatik, 2013

[2] Hirsch, H.G., Kremer, F.: Recognition of Noisy Speech by Starting the Likelihood Calculation at Voiced Segments, ITG Fachtagung Sprachkommunikation, 2014

[3] FaNT – Filtering and Noise Adding Tool, verfügbar im Download Bereich unter http://dnt.kr.hs-niederrhein.de

[4] Aurora Sprachdatenbasen, URL: http://aurora.hs-niederrhein.de, Datenbasen werden vertrieben von ELRA: http://www.elda.org

[5] Kolossa, D., Zeiler, S., Saeidi, R., Astudillo, R.F.: Noise Adaptive LDA: A New Approach for Speech Recognition under Observation Uncertainity, IEEE Signal Processing Letters, Vol. 20, 2013