Automatische Sprecherverifikation basierend auf Stimmgrundfrequenz-Merkmalen mittels Hauptkomponentenanalyse

Timo Becker, Wolfgang Kreuzer
Österreichische Akademie der Wissenschaften, Institut für Schallforschung

Einleitung

In der automatischen Sprechererkennung haben sich bisher auf spektralen/cepstralen Merkmalen basierende Systeme bewährt. In den vergangenen Jahren sind zusätzliche Merkmale hinzugekommen, die auch mit traditionellen Merkmalen kombiniert werden, um die Performanz der Systeme weiter zu verbessern [3]. Die Stimmgrundfrequenz (F0) hat den Vorteil, dass sie nicht so stark von Unterschieden des Übertragungskanals beeinflusst wird. Kinoshita et al. [5] [6] haben gezeigt, dass auch nur auf diesen Merkmalen basierend eine Sprecherverifikation möglich ist. Ein Hauptproblem hierbei ist jedoch die Korrelation der Merkmale, die durch multivariate Modellierung in einem Bayes-Framework kompensiert wurde. Hierfür sind jedoch lange Signaldauern notwendig, da die intra-Sprecher-Variabilität durch Schätzung von Kovarianzmatrizen erfasst werden muss. In diesem Artikel wird ein Verfahren präsentiert, dass eine Verbesserung der Gleichfehlerrate dadurch erreicht, dass die Merkmalsdimensionen durch Hauptkomponentenanalyse optimiert werden.

Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (im Englischen Principal Component Analysis, PCA genannt) ermöglicht es, Muster innerhalb eines Datensatzes zu erkennen, und zusätzlich die Menge an Daten auf sinnvolle Weise zu reduzieren. Dazu wird die Kovarianzmatrix $C_{X_{\text{alt}}}$ eines multidimensionalen Datensatzes X_{alt} betrachtet, bei dem in jeder Spalte der Mittelwert über alle Samples abgezogen wurde. Nach einer Eigenwertzerlegung

$$C_{X_{\text{alt}}} = EDE^T \tag{1}$$

mit Eigenvektoren E und der Diagonalmatrix D, die die Eigenwerte enthält, erhält man mittels

$$X_{\text{neu}}^T = E^T X_{\text{alt}}^T \tag{2}$$

eine neue Basis X_{neu} aus Linearkombinationen von Merkmalsvektoren. Aufgrund der Tatsache, dass für die neue Kovarianzmatrix

$$C_{X_{\text{neu}}} = X_{\text{neu}}^T X_{\text{neu}} = E^T X_{\text{alt}}^T X_{\text{alt}} E = E^T EDE^T E = D \tag{3}$$

gilt, ist es ersichtlich, dass X_{neu} dekorreliert ist. Darüber hinaus ist es möglich den Datensatz zu reduzieren, indem Basisvektoren, die zu kleinen Eigenwerten gehören (d.h. Merkmalkombinationen, die keinen grossen Einfluss auf den Datensatz haben) einfach weggelassen werden, da sie zur Erklärung der Varianz wenig beitragen.

Stimmgrundfrequenzmerkmale

Die Stimmgrundfrequenz (F0) wurde mit Hilfe des F0-Extraktionsverfahren von STx [1] verwendet, das auf der Autokorrelation basiert [2]. Für 100 männliche Sprecher des Pool2010-Korpus (Studioaufnahmen, gelesen) [4] wurden die Aufnahmen halbiert und jeweils für Trainingsmenge (9,6 Sekunden durchschnittliche Dauer des stimmhaften Sprachsignals) und Testmenge (9,4 Sekunden durchschnittliche Dauer des stimmhaften Sprachsignals) verwendet. Folgende Merkmale wurden berechnet:

F0_{mean}	arithmetisches Mittel
F0_{sd}	Standardabweichung
F0_{VarCo}	Variationskoeffizient [4]
F0_{LE}	$\text{F0}_{\text{mean}} - 1{,}43 \cdot \text{F0}_{\text{sd}}$ [7]
F0_{Skew}	Skewness (Schiefe)
$\text{F0}_{\text{Kurtosis}}$	Kurtosis (Exzess)

Die Merkmale wurden einzeln und in Kombination als 6-dimensionaler Merkmalsvektor untersucht (im Folgenden kurz als *Vektor* bezeichnet). Um Korrelationen zu kompensieren, wurde anhand von 287 Sprechern des TIMIT-Korpus (Studioaufnahmen, gelesen, 14,8 Sekunden durchschnittliche Dauer des stimmhaften Sprachsignals) eine Hauptkomponentenanalyse durchgeführt, und die Basistransformationen verwendet, um die Merkmalsvektoren der Trainings- und Testaufnahmen auf die neue Basis abzubilden [9]. Dies geschieht unter der Annahme, dass die Basen unter vergleichbaren Bedingungen ähnlich sind. Tabelle 1 zeigt die erklärte Varianz der Basiskomponenten. Die resultierenden dekorrelierten Merkmale PC1, ..., PC6 wurden durch schrittweise Hinzunahme in den Merkmalsvektor untersucht.

Tabelle 1: Erklärte Varianz der Basiskomponenten in Prozent

PC1	PC2	PC3	PC4	PC5	PC6
93,48	5,77	0,71	0,04	0,00	0,00

Verifikationssystem

Als Unhähnlichkeitsmaß dient die Euklidsche Distanz. Die Performanz des Sprechererkennungssystem wurde mittels Gleichfehlerrate (GFR) bestimmt (siehe Tabelle 2) und in einem Detection Error Tradeoff Plot (DET Plot) [8] dargestellt (siehe Abb. 1).

Tabelle 2: Gleichfehlerraten (GFR)

	GFR
$F0_{mean}$	0,20
$F0_{sd}$	0,32
$F0_{VarCo}$	0,33
$F0_{LE}$	0,23
$F0_{Skew}$	0,40
$F0_{Kurtosis}$	0,41
Vektor	0,15
PC1	0,15
PC1+PC2	0,10
PC1+PC2+PC3	0,10
PC1+PC2+PC3+PC4	0,10
PC1+PC2+PC3+PC4+PC5	0,10
PC1+PC2+PC3+PC4+PC5+PC6	0,10

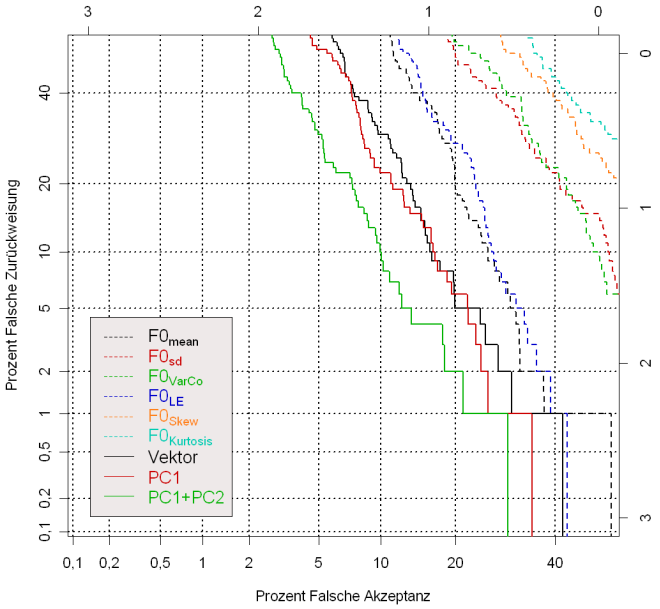

Abbildung 1: Detection Error Tradeoff Plot (Basistransformation nur bis maximal 2 Dimensionen)

Ergebnisse

Am besten diskriminiert das Merkmal $F0_{mean}$ mit einer GFR von 20 %. Die Zusammenfassung aller sechs Merkmale in einem Merkmalsvektor senkt die GFR weiter auf 15 %. Durch Anwendung der Hauptkompontenanalyse und Verwendung der ersten Komponente wird das gleiche Ergebnis wie mit dem sechsdimensionalen Merkmalsvektor erzielt. Die Verwendung von zwei und mehr Komponenten senkt die Fehlerrate weiter auf 10 %. Durch Anwendung der Hauptkomponentenanalyse kann somit neben einer Dimensionsreduktion eine Verbesserung der GFR erreicht werden. Vergleichbare Ergebnisse wurden von Kinoshita [6] erreicht, jedoch nur mit Signaldauern von mehr als 10 Minuten Länge.

Literatur

[1] `www.kfs.oeaw.ac.at`.

[2] Boersma, Paul: *Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound.* IFA Proceedings, 17:97–110, 1993.

[3] Farrús, Mireia, Javier hernando und Pascual Ejarque: *Jitter and Shimmer Measurements for Speaker Recognition.* In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Seiten 778–781, 2007.

[4] Jessen, Michael, Olaf Köster und Stefan Gfroerer: *Influence of vocal effort on average and variability of fundamental frequency.* Speech, Language and the Law, 12(2):174–213, 2005.

[5] Kinoshita, Yuko: *Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0.* Speech, Language and the Law, 12(2):235–254, 2005.

[6] Kinoshita, Yuko, Shunichi Ishihara und Phil Rose: *Beyond the long-term mean: Multivariate likelihood ratio-based FSR using F0 distribution parameters.* In: *Proceedings of the IAFPA*, Seite 15, 2007.

[7] Lindh, Jonas und Anders Eriksson: *Robustness of Long Time Measures of Fundamental Frequency.* In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Seiten 2025–2028, 2007.

[8] Martin, Alvin, George Doddington, Terri Kamm, Mark Ordowski und Mark Przybocki: *The DET Curve in Assessment of Detection Task Performance.* In: *Proc. Eurospeech '97*, Seiten 1895–1898, 1997.

[9] R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.