Stimmmerkmale für die Sprecherverifikation

Martin Gämperle

Semesterarbeit SA06.27

Sommersemester 2006
Abgabedatum: 07. Juli 2006

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: Dr. B. Pfister und M. Gerber

Verantwortlicher: Prof. Dr. L. Thiele

Inhaltsverzeichnis

Zusammenfassung

Die Aufgabe dieser Arbeit war, zu testen, ob die Sprecherverifikation mit den üblich dafür verwendeten MFCCs (Mel-Frequency Cepstral Coefficients, siehe [1] und [2]) mit weiteren Merkmalen, wie z.B. der Grundfrequenz und deren Modulation, verbessert werden könnte.

Getestet wurden die Grundfrequenz und deren ersten beiden Ableitungen einzeln, in einer gewichteten Kombination und schliesslich zusammen mit den MFCC-Merkmalen (MFCCs, erste und zweite Ableitung). Der Versuch mit Jitter und Shimmer (siehe [3] und [4]), d.h. mit der Modulation der Periodendauer bzw. der Amplitude, scheiterte bereits beim Extrahieren dieser Merkmale. Im Rahmen dieser Arbeit war es nicht möglich, dafür eine Methode zu finden, die auf die verwendeten Telefon-Signale anwendbar gewesen wäre.

Die Merkmale wurden aus einer Sammlung von Telefon-Sprachsignalen extrahiert und statistisch ausgewertet. Die einzelnen Grundfrequenz-Merkmale und deren Kombination haben zwar sprecherunterscheidendes Potential, sind aber weniger aussagekräftig als die MFCC-Merkmale. Zusammen mit den MFCC-Merkmalen konnte eine geringe Verbesserung erzielt werden, im Vergleich zur Kombination der MFCC-Merkmale alleine.

1 Einführung

1.1 Sprecherverifikation

Wir Menschen sind in der Lage, Personen anhand ihrer Stimmen sofort zu erkennen, selbst wenn das Sprachsignal über einen schlechten Kanal, wie z.B. das Telefon, übertragen wird.

Die menschliche Stimme enthält also Merkmale, die es uns erlauben, verschiedene Sprecher zu erkennen bzw. sie zu unterscheiden. Die Aufgabe der Sprecherverifikation ist es, solche Stimmmerkmale zu extrahieren und mit deren Hilfe festzustellen, ob zwei Sprachsignale vom gleichen Sprecher gesprochen worden sind oder nicht.

In dieser Arbeit wurde ein textabhängiges Verfahren verwendet, d.h. der Inhalt des Sprachsignals war bekannt und es wurden Signalsegmente mit den selben gesprochenen Wörtern oder Sätzen verglichen.

1.2 Aufgabenstellung

Die Aufgabe dieser Arbeit war, weitere Stimmmerkmale zu finden, die eventuell das Resultat der Sprecherverifikation mit den üblich verwendeten MFCC-Merkmalen (Mel-Frequency Cepstral Coefficients, siehe [1] und [2]) verbessern könnten. Die genaue Aufgabenstellung vom Institut ist im Anhang untergebracht.

Die Merkmalsextraktion für die Sprecherverifikation sollte auch mit qualitativ schlechten Sprachsignalen funktionieren, da einige Anwendungen, z.B. in der Forensik, über Telefone übermittelte Sprachsignale verwenden.

2 Wahl der Stimmmerkmale

2.1 Grundfrequenz und deren Ableitungen

Da stimmhafte Segmente im Sprachsignal quasi-periodisch sind, weisen diese eine Grundwelle und Oberwellen auf. Die Frequenz der Grundwelle wird als Grundfrequenz oder F_0 bezeichnet. Die Grundfrequenz liefert die Tonhöhe, in der der Sprecher spricht. Bei Männerstimmen bewegt sie sich ungefähr im Bereich von 60 Hz bis 120 Hz, bei Frauenstimmen etwa 120 Hz bis 200 Hz (siehe [4]). Da die MFCCs die Enveloppe des Kurzzeitspektrums repräsentieren, ist in ihnen die Grundfrequenz nicht zu erkennen. Darum ist zu vermuten, dass die Grundfrequenz als Stimmmerkmal zusätzliche Information über den Sprecher liefern könnte.

Gewisse Sprecher neigen vielleicht dazu, zu Beginn jedes stimmhaften Segments die Stim-

me zu heben und gegen Ende wieder fallen zu lassen, andere eventuell gerade umgekehrt. Oder die Stimme steigt ständig an oder fällt immer wieder ab. Deshalb macht es Sinn, auch die ersten beiden zeitlichen Ableitungen der Grundfrequenz zu untersuchen. Die erste Ableitung gibt an, wie schnell die Grundfrequenz ansteigt bzw. abfällt. Mit der zweiten Ableitung erhält man Informationen darüber, wie stark und in welche Richtung der Verlauf der Grundfrequenz gebogen ist.

2.2 Jitter und Shimmer

In der Literatur über Stimmmerkmale (z.B. in [3] und [4]) werden oft Jitter und Shimmer erwähnt. Jitter ist ein Mass für die Variabilität der Periodendauer, Shimmer für die Änderung der Amplitude von Periode zu Periode. Es gibt dazu eine ganze Reihe von unterschiedlichen mathematischen Umsetzungen. In [4] werden z.B. der *Local Jitter* und der *Local Shimmer* wie folgt definiert:

$$jitter_{loc} = \frac{100}{(N-1)\overline{T}} \sum_{p=2}^{N} |T(p) - T(p-1)| \qquad [\%] \tag{1}$$

$$shimmer_{loc} = \frac{100}{(N-1)\overline{A}} \sum_{p=2}^{N} |A(p) - A(p-1)| \qquad [\%] \tag{2}$$

N ist die Anzahl von Perioden, $T(p)$ die Periodendauer und $A(p)$ die Amplitude der Periode p. $\overline{T}$ und $\overline{A}$ stehen für die durchschnittliche Periodendauer bzw. Amplitude.

Für beide dieser Merkmale ist es notwendig die Periodendauer bzw. die Amplitude jeder Periode exakt zu bestimmen. Da jedoch die untersuchten Sprachsignale über Telefone übertragen worden sind, enthalten diese nur noch Frequenzkomponenten zwischen etwa 300 Hz und 3400 Hz (siehe [1]), die Grundwelle, v.a. bei Männerstimmen, ist also nicht mehr vorhanden, ausserdem enthalten die Signale relativ viel Rauschen und Verzerrungen. Diese Faktoren machten die hier für die Jitter- bzw. Shimmer-Analyse angewandte Detektion von Nulldurchgängen oder Maxima bzw. Minima praktisch unmöglich. Die sehr kleinen Schwankungen gingen im Rauschen unter. Eventuelle andere, komplexere Methoden, hätten den Rahmen dieser Arbeit gesprengt, Jitter und Shimmer wurden deshalb nicht weiter untersucht.

3 Extraktion von Stimmmerkmalen

3.1 Grundfrequenz

Für die Extraktion der Grundfrequenz stand bereits eine am Institut entwickelte Funktion zur Verfügung. Sie berechnet zuerst das Cepstrogramm und mit dessen Hilfe den ungefähren Verlauf der Grundfrequenz. Anschliessend wird nach dieser approximierten Grundfrequenz oder,

wenn die Grundwelle nicht mehr vorhanden ist, nach deren harmonischen Oberwellen[1] im Spektrogramm gesucht. Werden sie nicht gefunden, so wird das entsprechende Frame[2] als nicht stimmhaft zurückgegeben. Wenn sie gefunden werden, verfolgt die Funktion die Hügelzüge im Spektrogramm und sucht deren Enden. So können meist weitere stimmhafte Frames gefunden werden. Für stimmlose Frames wird die Grundfrequenz linear zwischen dem letzten stimmhaften davor und dem ersten stimmhaften danach interpoliert.

Die Abbildung 1 zeigt ein Beispiel eines Signalsegments und das dazugehörige Spektrogramm und Cepstrogramm. Die Quasi-Periodizität von stimmhaften Lauten ist an den dunklen Konturen im Cepstrogramm zu erkennen. Im Spektrogramm sind die harmonischen Oberwellen als dunkles Linienmuster gut ersichtlich.

[1] Der Abstand der harmonischen Oberwellen entspricht der Grundfrequenz (siehe [1]).

[2] Jedes Frame entspricht einer Position des Analysefensters. Für jedes Frame wird ein Merkmalsvektor berechnet.

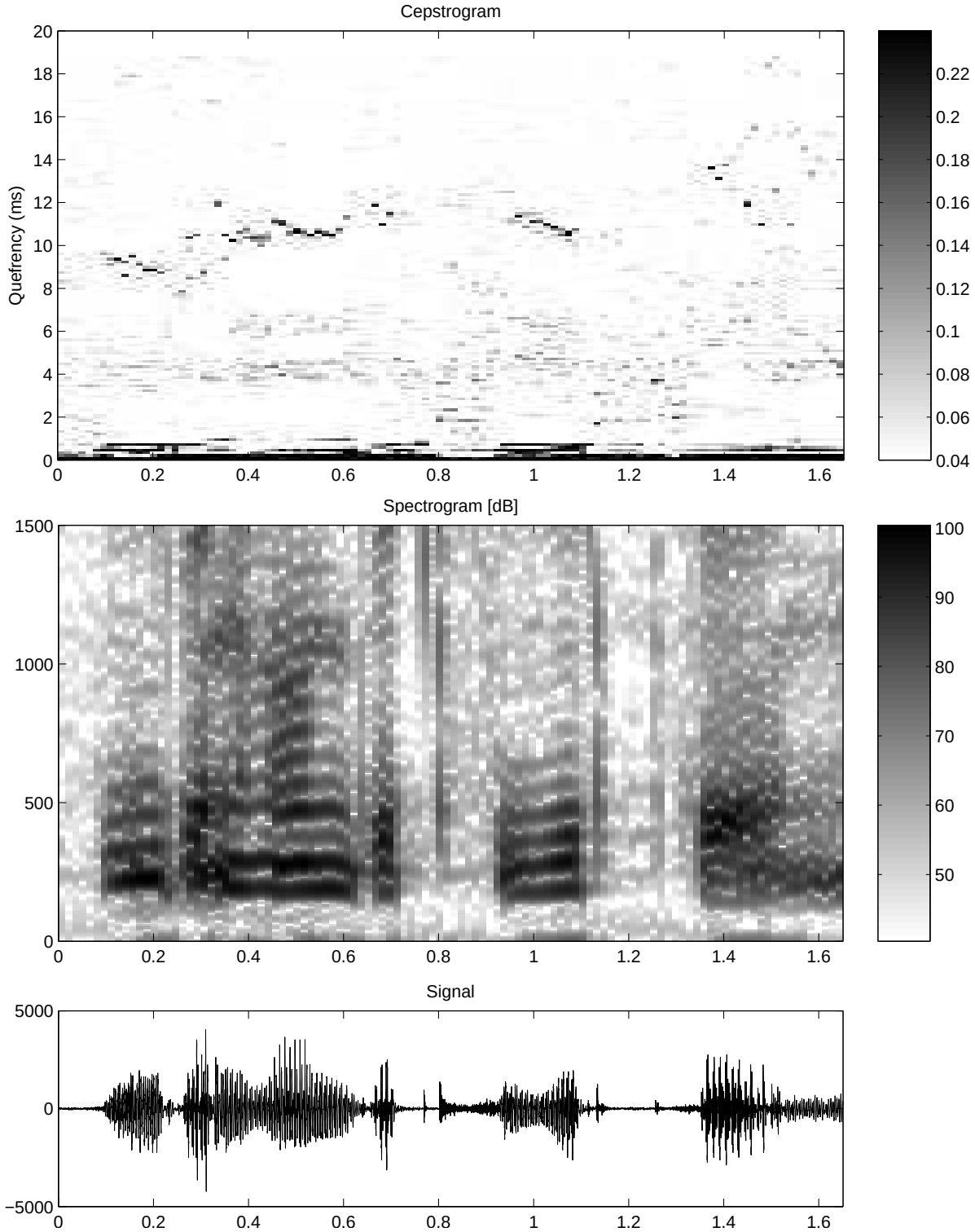

Abbildung 1: *Cepstrogramm, Spektrogramm und Signal, in dem "siebenhundertvierzehn" gesprochen worden ist. Mit Hilfe des Spektrogramms wird der mit dem Cepstrogramm berechnete, ungefähre Verlauf der Grundfrequenz verifiziert, bzw. durch verfolgen der Hügelzüge (dunkle Stellen) erweitert.*

3.2 Ableitungen der Grundfrequenz

Die Differenz von jeweils zwei zeitlich aufeinander folgenden F_0-Werten liefert eine sehr verrauschte Approximation der ersten Ableitung. Deshalb wird hier eine Methode verwendet, mit der die geschätzte Ableitung geglättet wird. Auf diese Weise wird der Rauschanteil auf Kosten der zeitlichen Auflösung verringert. Der Verlauf der Grundfrequenz wird dazu in einem Zeitfenster mit einer Breite von $2L+1$ Frames durch ein Polynom approximiert. Anschliessend wird dieses Polynom abgeleitet. Für die erste Ableitung der Grundfrequenz zur Zeit t liefert diese Approximation die Formel (siehe [1])

$$\Delta F_0(t) = \frac{\sum_{l=-L}^{L} l\, F_0(t+l)}{\sum_{l=-L}^{L} l^2} \qquad \text{[Hz/Frame]} \tag{3}$$

Die Schätzung der zweiten Ableitung wird mit dem wiederholten Anwenden der Formel (3) auf die erste Ableitung (statt auf F_0) erzeugt.

An den Rändern des abzuleitenden Signalsegments stehen zu wenig Frames zur Verfügung, daher muss dort die Breite des Zeitfensters entsprechend verringert werden. Dadurch sind die Werte an den Rändern gezwungenermassen etwas verrauschter als die anderen.

3.3 Ein Beispiel

In der Abbildung 2 ist ein Beispiel eines Segments in dem ”siebenhundertvierzehn” gesprochen worden ist. Sie zeigt das Signal, den Grundfrequenzverlauf, deren ersten zwei geglätteten Ableitungen und den Gewichtsvektor, der stimmhafte Frames mit 1 und stimmlose mit 0 gewichtet.

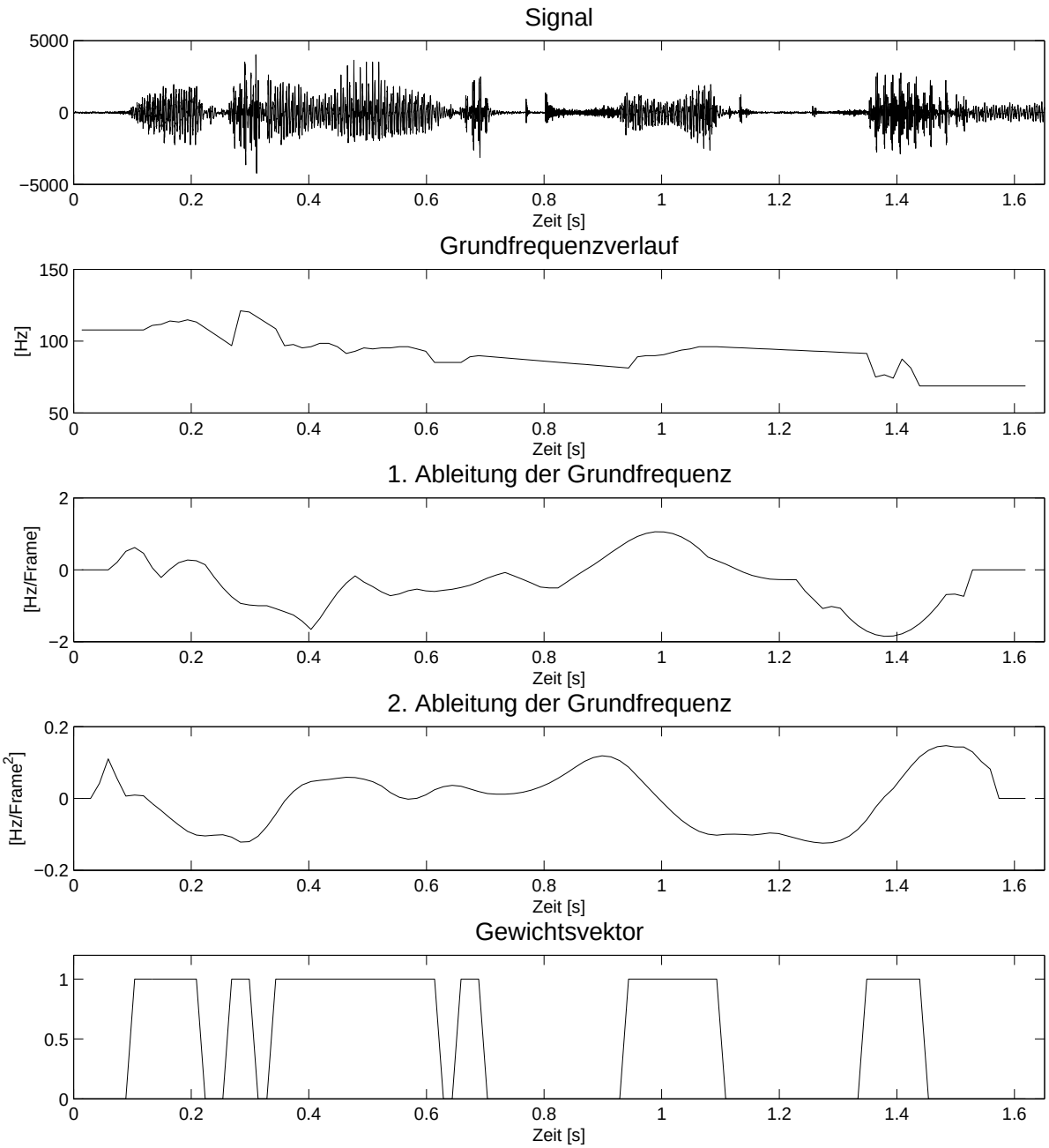

Abbildung 2: *Signal, Grundfrequenzverlauf, erste und zweite Ableitung der Grundfrequenz und Gewichtsvektor eines Signalsegments, in dem ”siebenhundertvierzehn” gesprochen worden ist. Im Gewichtsvektor besitzen die stimmhaften Frames den Wert 1 und die stimmlosen 0.*

4 Testen der Stimmmerkmale

4.1 Verwendete Sprachsignale

Eine Sammlung von aufgezeichneten Telefon-Sprachsignalen, inklusive Informationen über den gesprochenen Inhalt und die Identität des Sprechers, war bereits vorhanden. Jedes Signal enthielt dieselben 15 gesprochenen Zahlwörter von dreistelligen Zahlen und dieselben 5 kurzen Sätze. Allerdings waren sie von verschiedenen, männlichen Sprechern bzw. über unterschiedliche Telefone aufgezeichnet worden. Die digitalen Signale standen mit einer Abtastrate von 8000 Hz zur Verfügung.

4.2 Parameterwahl für die Merkmalsextraktion

Für die F_0-Analyse wurde ein Hamming-Fenster mit der Breite von 300 Samples und einer Verschiebung um jeweils 120 Samples verwendet. Für die Approximation der Ableitungen konnten mit einer Fensterbreite von 17 Frames ($L = 8$) die besten Resultate erzielt werden. Diese Fensterbreite ist, im Vergleich zu anderen Delta-Merkmalen, relativ gross. Die Grundfrequenz scheint also weniger rasch zu variieren, als andere Merkmale (z.B. MFCCs).

4.3 Paarbildung

Für die Auswertung wurden zuerst Paare von unterschiedlichen Signalen, in denen dasselbe gesprochen wurde, gebildet. Einerseits solche Paare mit Signalen vom gleichen Sprecher (Klasse *self*) und andererseits solche von unterschiedlichen Sprechern (Klasse *cross*). Insgesamt wurden 154 Paare der Klasse *self* und 477 Paare der Klasse *cross* verwendet.

4.4 Bestimmen der Scores für ein Merkmal

Weil hier ein textabhängiges Experiment durchgeführt wurde und daher die gesprochenen Inhalte bereits bekannt waren, mussten nur in jedem Signalpaar jeweils die entsprechenden Segmente mit gleichem Inhalt (die bekannten Zahlwörter bzw. kurzen Sätze) aus beiden Signalen miteinander verglichen werden. Dazu wurden, unter Verwendung des DTW-Algorithmus (Dynamic Time Warping Algorithm, siehe [1]), die entsprechenden Frames zeitlich angepasst und daraus Framepaare gebildet. Für jedes Framepaar berechnete man anschliessend die absolute Differenz zwischen den beiden Merkmalswerten und negierte diese, damit eine kleinere Zahl auch einer kleineren Wahrscheinlichkeit entspricht, dass die beiden Frames aus dem Paar vom gleichen Sprecher stammen. Diese negierten Distanzen wurden als *Scores* bezeichnet.

4.5 Kombinieren mehrerer Merkmale

Da die verschiedenen Merkmale auch unterschiedliche Einheiten besitzen, mussten, für die Kombinationen mehrerer Merkmale, die Scores der einzelnen Merkmale so normiert werden,

dass die optimale Klassierungsgrenze etwa bei 0 lag und die Varianzen der Verteilungen ungefähr gleich gross waren. Nach dieser Normierung wurde für jedes Framepaar der gewichtete Mittelwert aus den Scores aller Merkmale gebildet. Die Gewichte wurden so gewählt, dass mit den daraus resultierenden Scores die bestmögliche Unterscheidung der beiden Klassen erzielt werden konnte.

4.6 Statistische Auswertung

Ab hier wurden zwei Varianten untersucht: Bei der ersten wurden die Framepaar-Scores für Paare von Signalabschnitten mit einer Länge von 60 Frames, das entspricht ungefähr einer Sekunde, erneut gemittelt. Bei der zweiten Variante wurde der Durchschnitt aus den Framepaar-Scores über die ganzen Signalpaare gebildet. So blieb schliesslich nur noch ein gemittelter Score pro Signalabschnittpaar bzw. Signalpaar übrig.

Bei der Mittelung der Scores aus den F_0-Merkmalen wurden die stimmlosen Frames nicht betrachtet. Enthielten also die oben erwähnten 60 Frames sehr viele stimmlose, so trugen nur die wenigen übrigen stimmhaften Frames zum Mittelwert des Signalabschnitts bei.

Nach dem Bestimmen der beiden Häufigkeitsverteilungen der gemittelten Scores aller Signalabschnittpaare bzw. Signalpaare der Klasse *self* und der Klasse *cross* wurden sie aufsummiert, diejenige der Klasse *cross* noch von 1 subtrahiert, und schliesslich geplottet (siehe Abbildungen 3 bis 10 und [5]).

Wenn sich die Kurven der beiden aufsummierten Verteilungen nicht schneiden, dann lassen sich alle untersuchten Paare problemlos der richtigen Klasse zuordnen. Schneiden sich die beiden Kurven, so gibt es einen Bereich, in dem beide Wahrscheinlichkeiten grösser als 0 sind. Paare deren Scores in diesem Bereich liegen, lassen sich nicht mit absoluter Sicherheit der richtigen Klasse zuordnen. Ausserdem kann aus dem Plot abgelesen werden, welche Scores am häufigsten auftraten, nämlich diejenigen an den Stellen, wo die aufsummierten Verteilungen jeweils die grösste Steigung aufweisen. Für eine gute Trennung, sollten diese Maxima möglichst weit voneinander entfernt sein.

4.7 Fisher Kriterium

Um einen Zahlenwert zu erhalten, der die Qualität der Merkmale widerspiegelt, wurde das Fisher Kriterium F mit der Formel (4) berechnet.

$$F = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2} \tag{4}$$

Wobei μ_i der Mittelwert und σ_i die Standardabweichung der Verteilung der Klasse i ist.

Das Fisher Kriterium gilt als Mass für die Distanz zwischen den Verteilungen der beiden Klassen *self* und *cross*. Je höher das Fisher Kriterium desto deutlicher lässt sich also entscheiden, ob derselbe Sprecher gesprochen hat oder nicht.

5 Resultate

5.1 Die einzelnen F_0-Merkmale

Die Abbildungen 3 und 4 zeigen die aufsummierten Wahrscheinlichkeitsverteilungen resultierend aus der Grundfrequenz bzw. deren ersten und zweiten Ableitung jeweils separat ausgewertet.

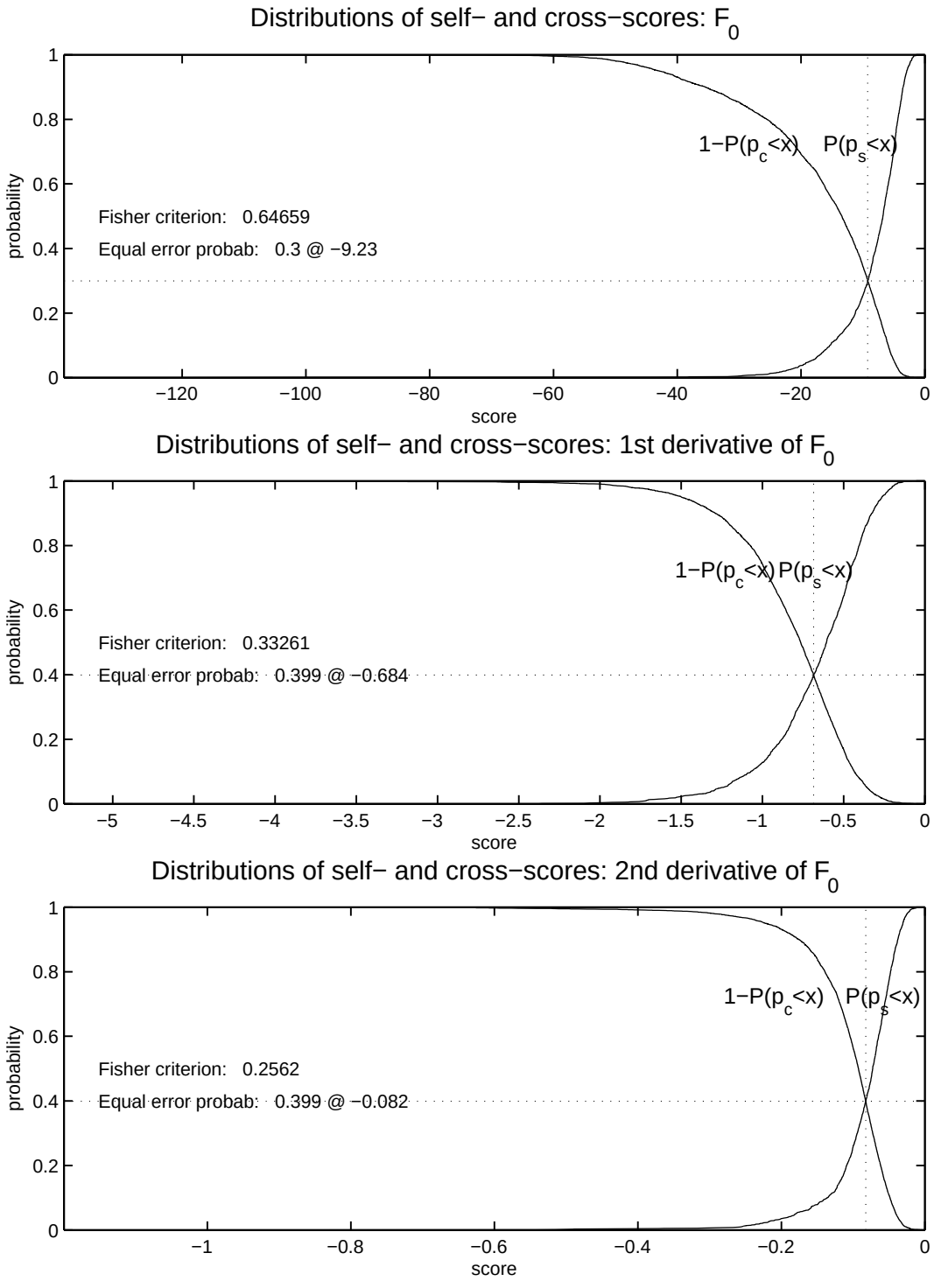

Abbildung 3: *Scoreverteilungen der Merkmale F_0 (oben), deren ersten Ableitung (Mitte) und deren zweiten Ableitung (unten) gemittelt über Signalabschnitte mit einer Länge von ungefähr einer Sekunde.*

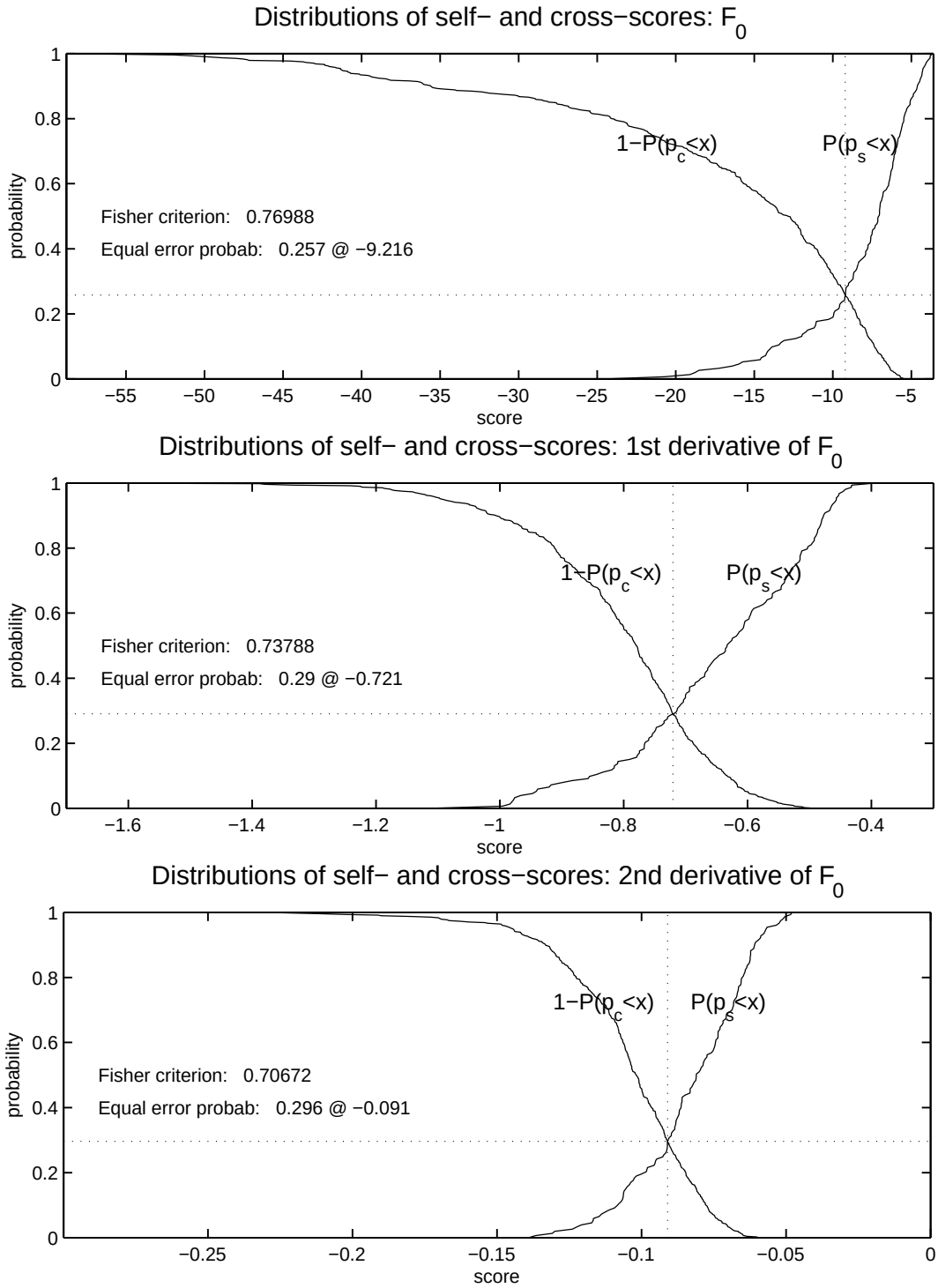

Abbildung 4: *Scoreverteilungen der Merkmale F_0 (oben), deren ersten Ableitung (Mitte) und deren zweiten Ableitung (unten) gemittelt über die ganzen Signale.*

Gemittelt über die Signalabschnitte mit einer Länge von etwa einer Sekunde ergab der Versuch für die Grundfrequenz als alleiniges Merkmal ein Fisher Kriterium von 0.6466, für die erste Ableitung 0.3326 und für die zweite Ableitung 0.2562. Die dazugehörigen Equal Error Probabilities waren 0.300, 0.399 und 0.399. Sämtliche dieser drei Plots sind stark asymmetrisch und die Stellen, an denen die Steigung und somit auch die Wahrscheinlichkeit am grössten sind, liegen überall sehr dicht beieinander. Bei der Mittelung über die ganzen Signale, lieferte die Grundfrequenz ein Fisher Kriterium von 0.7699, die erste Ableitung 0.7379 und die zweite Ableitung 0.7067. Die entsprechenden Equal Error Probabilities waren 0.257, 0.290 und 0.296. Der Plot resultierend aus der Grundfrequenz, weist eine starke Asymmetrie auf, folglich spricht ein bestimmter Sprecher immer etwa in der gleichen Tonhöhe, in der Klasse *cross* aber sprechen zwar viele in einer ähnlichen Stimmlage, andere Sprecher jedoch liegen weit auseinander. Die Stellen, an denen die Steigung am grössten sind, liegen überall, besonders aber beim Plot der Grundfrequenz, sehr nahe beieinander.

Die Resultate der gewichteten Kombinationen der drei F_0-Merkmale sind in den Abbildungen 5 und 6 zu sehen. Bei der Mittelung über die Signalabschnitte von einer Sekunde erwies sich eine Gewichtsverteilung zwischen den einzelnen Merkmalen von 0.7 für die Grundfrequenz, 0.2 für die erste und 0.1 für die zweite Ableitung als optimal. Die beste Gewichtsverteilung bei der Mittelung über die ganzen Signale war 0.40 für die Grundfrequenz, 0.35 für die erste und 0.25 für die zweite Ableitung.

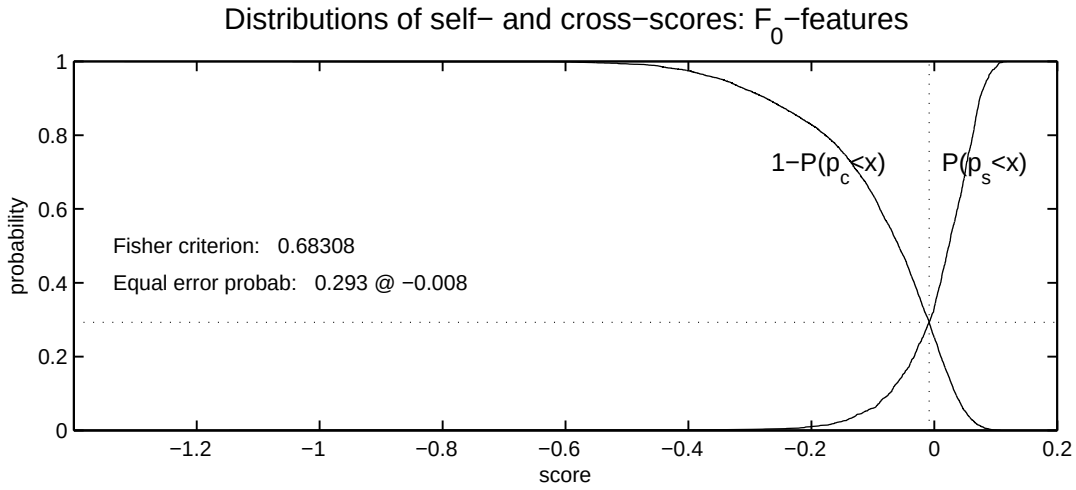

Abbildung 5: *Scoreverteilungen der Kombination aller F_0-Merkmale mit der Mittelung über Signalabschnitte mit einer Länge von etwa einer Sekunde.*

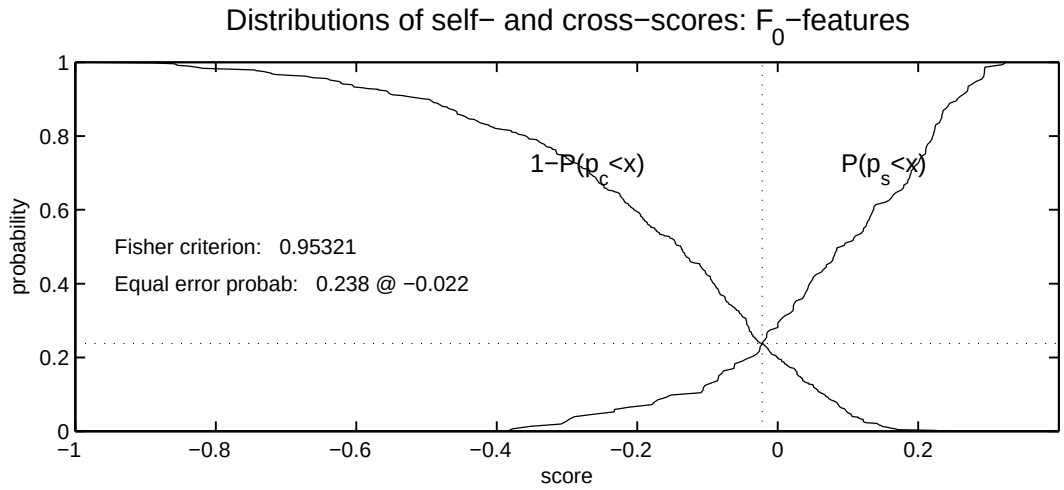

Abbildung 6: *Scoreverteilungen der Kombination aller F_0-Merkmale mit der Mittelung über die ganzen Signale.*

Diese Kombinationen lieferten ein Fisher Kriterium von 0.6831 und eine Equal Error Probability von 0.293 mit der Mittelung über die Signalabschnitte, 0.9532 und 0.238 bei der Mittelung über die ganzen Signale. Der Abstand der Stellen mit den grössten Wahrscheinlichkeiten ist auch hier relativ klein, v.a. bei der Mittelung über die Signalabschnitte.

5.2 MFCC zum Vergleich

In den Abbildungen 7 und 8 sind die Resultate der gewichteten Kombinationen der mit Neuronalen Netzen berechneten Wahrscheinlichkeiten aus den MFCC-Merkmalen zu sehen (siehe [6]). Die Gewichte wurden optimiert und folgendermassen verteilt: 0.5 für die MFCC, 0.3 für die erste und 0.2 für die zweite Ableitung.

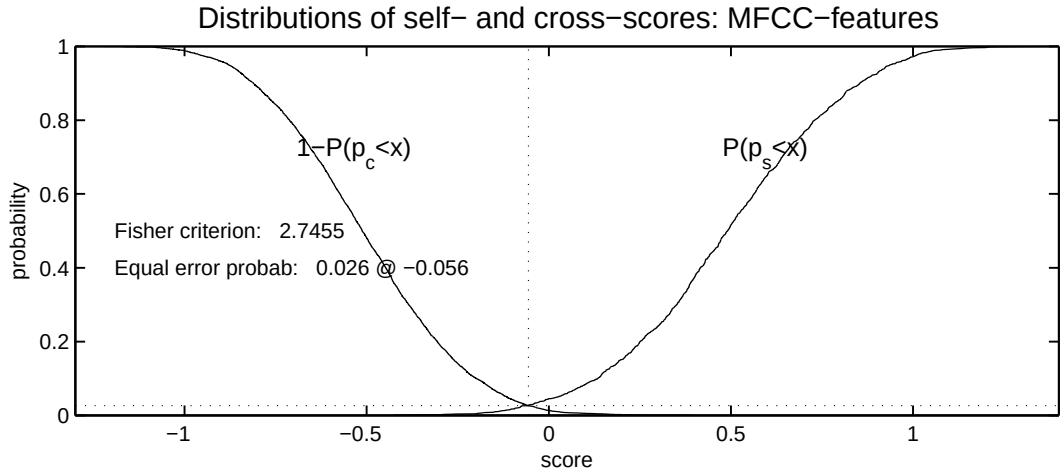

Abbildung 7: *Scoreverteilungen der gewichteten Kombination aus dem MFCC-Merkmal und dessen ersten beiden Ableitungen gemittelt über die kurzen Signalabschnitte.*

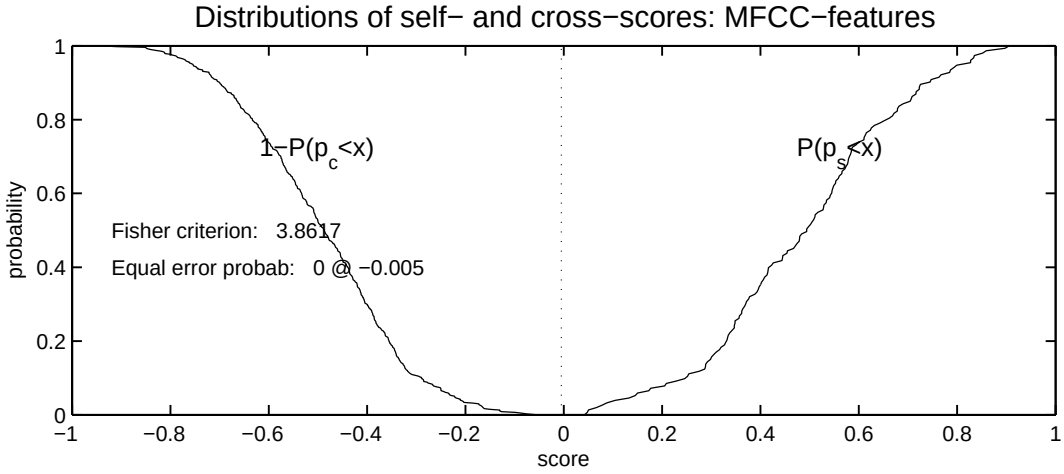

Abbildung 8: *Scoreverteilungen der gewichteten Kombination aus dem MFCC-Merkmal und dessen ersten beiden Ableitungen gemittelt über die ganzen Signale.*

Mit der Mittelung über die kurzen Signalabschnitte ergab sich ein Fisher Kriterium von 2.7455 und eine Equal Error Probability von 0.026. Über die ganzen Signale gemittelt lieferten die MFCC-Merkmale ein Fisher Kriterium von 3.8617. Die Verteilungen sind bei beiden Varianten praktisch symmetrisch. Bei der Mittelung über die ganzen Signale entstand kein Schnittpunkt zwischen den Verteilungskurven, d.h. die beiden Klassen konnten vollständig separiert werden.

5.3 Kombination der F_0- und MFCC-Merkmale

Um festzustellen, ob die F_0-Merkmale das Resultat der MFCC-Merkmale zu verbessern vermögen, wurde für jede Variante eine weitere gewichtete Kombination aller MFCC- und F_0-Merkmale getestet. Ein Gewichtsverhältnis zwischen den MFCC- und den F_0-Merkmalen von 0.66 zu 0.34 bei der Mittelung über die Signalabschnitte und von 0.73 zu 0.27 bei der Mittelung

über die ganzen Signale erwies sich als optimal. Die Ergebnisse dieser Kombinationen zeigen die Abbildungen 9 und 10.

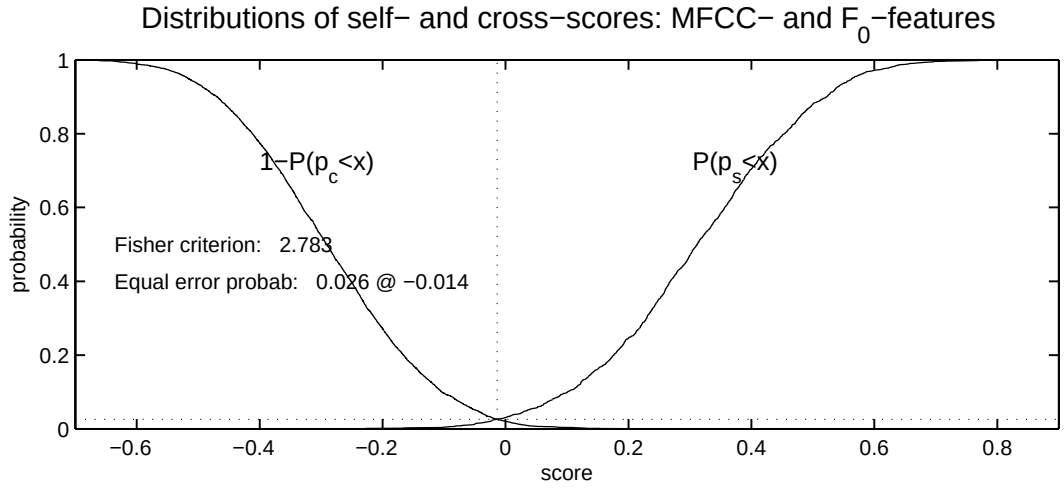

Abbildung 9: *Scoreverteilungen der Kombination aller Merkmale (MFCC, F_0 und jeweils die erste und zweite Ableitung) gemittelt über die kurzen Signalabschnitte.*

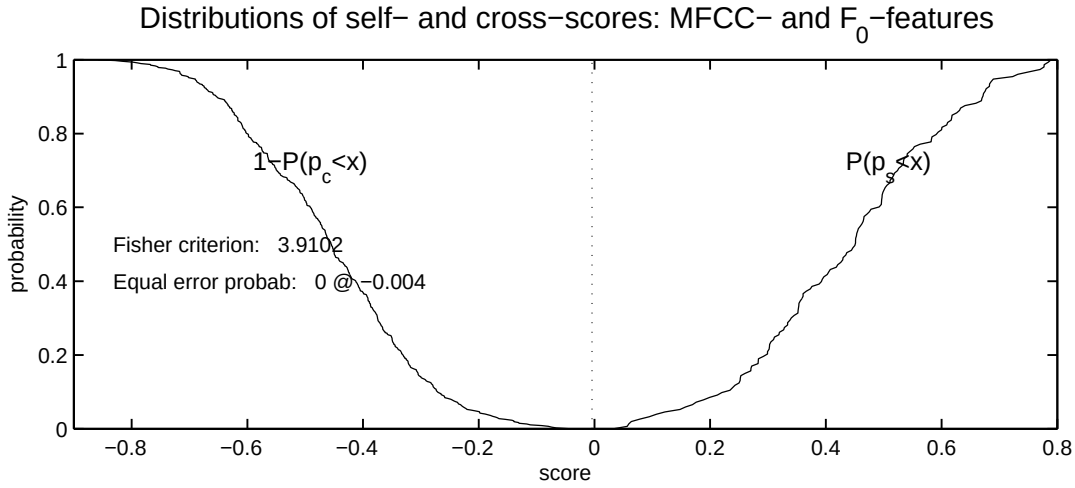

Abbildung 10: *Scoreverteilungen der Kombination aller Merkmale (MFCC, F_0 und jeweils die erste und zweite Ableitung) gemittelt über die ganzen Signale.*

Mit der Mittelung über die Signalabschnitte ergaben sämtliche sechs Merkmale kombiniert ein Fisher Kriterium von 2.783 und eine Equal Error Probability von 0.026. Gemittelt über die ganzen Signale, ergab die Kombination ein Fisher Kriterium von 3.9102. Auch hier sind die Kurven beider Plots beinahe symmetrisch. Die Kurven in Abbildung 10 weisen keinen Schnittpunkt auf, also war mit der Mittelung über die ganzen Signale die absolute Trennung der Klassen möglich.

6 Diskussion

Die Grundfrequenz lieferte mehr Information über den Sprecher als deren Ableitungen. Hier ist jedoch zu beachten, dass die zeitliche Auflösung mit jeder Ableitung abnimmt und damit auch die Qualität der Merkmalsextraktion. Das Resultat der Kombination aller F_0-Merkmale war besser, als die Ergebnisse der einzelnen Merkmale. Allerdings war es mit keinem dieser Merkmale oder der Kombination davon möglich, die Verteilungen der beiden Klassen vollständig zu trennen, was z.B. mit den MFCC-Merkmalen (gemittelt über die ganzen Signalpaare) gelang.

Die F_0-Merkmale und die MFCC-Merkmale zusammen vermochten das Resultat im Vergleich zur Kombination der MFCC-Merkmale alleine kaum relevant verbessern. Mit der Mittelung über die kurzen Signalabschnitte konnte das Fisher Kriterium nur um knapp 1.4 % von 2.7455 auf 2.7830, mit der Mittelung über die ganzen Signale um etwa 1.3 % von 3.8617 auf 3.9102 gesteigert werden.

An den Grenzen zwischen stimmhaften und stimmlosen Frames wurden für die Approximation der F_0-Ableitungen Werte verwendet, die im stimmlosen Bereich linear interpoliert wurden. So nehmen die Werte kurz vor stimmlosen Lauten Einfluss auf die Ableitung an der Stelle kurz nach den stimmlosen Lauten und umgekehrt. Allerdings sind diese Werte nur sehr ungenau. Würde man dies ändern, so könnten für kurze stimmhafte Segmente wiederum nur sehr verrauschte Approximationen der Ableitungen berechnet werden, weil dort die Fensterbreite verringert werden müsste.

7 Vorschläge für allfällige Folgearbeiten

Es ist zu vermuten, dass Jitter und Shimmer relativ viel über einen Sprecher aussagen würden. Der Versuch mit der Detektion von Nulldurchgängen, Maxima oder Minima schlug jedoch fehl. Es könnte nach weiteren, eventuell komplexeren Methoden gesucht werden, mit denen man Jitter und Shimmer aus Telefon-Sprachsignalen extrahieren könnte.

Relativ viele Parameter wurden in dieser Arbeit für jeweils jedes Teilproblem optimiert. Es ist jedoch durchaus möglich, dass z.B. bei der Kombination der Merkmale andere Parameter besser wären als jene, die für die einzelnen Merkmale optimal sind. Eine automatisierte Parameteroptimierung für das ganze Problem wäre daher eine Möglichkeit um bessere Resultate zu finden.

Da die Merkmale unterschiedliche Einheiten aufweisen, mussten die Differenzen normiert werden, bevor der Mittelwert über alle Merkmale pro Framepaar gebildet werden konnte. Diese Normierung gelang aufgrund der z.T. asymmetrischen Verteilungen nicht überall exakt. Die Verwendung eines Neuronalen Netzes würde diese ungenaue Normierung überflüssig machen

und deshalb vermutlich bessere Resultate liefern (siehe [6]).

Literatur

[1] B. Pfister and R. Beutler. Sprachverarbeitung I. *Vorlesungsskript, TIK, ETH Zürich*, 2005.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Gracía, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, pages 430–451, April 2004.

[3] I. R. Titze. Workshop on acoustic voice analysis. *National Center for Voice and Speech*, 1994.

[4] Ch. Müller. Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht. *Dissertation, Universität des Saarlandes*, 2005.

[5] B. Pfister. Personenidentifikation anhand der Stimme. *Kriminalistik*, April 2001.

[6] M. Gerber and B. Pfister. Quasi text-independent speaker verification with neural networks. *MLMI'05 Workshop, Edinburgh (United Kingdom)*, July 2005.

Anhang: Aufgabenstellung

Auf den nächsten Seiten folgt die Aufgabenstellung vom Institut.

Institut für
Technische Informatik und
Kommunikationsnetze

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

Sommersemester 2006

Semesterarbeitsaufgabenstellung

für

Herrn Martin Gämperle

Betreuer: M. Gerber ETZ D97.4
Stellvertreter: Dr. B. Pfister ETZ D97.6

Ausgabe: 3. April 2006
Abgabe: 7. Juli 2006

Stimmmerkmale für die Sprecher-Verifikation

Einleitung

Ein Sprachsignal enthält nicht nur Informationen über den gesprochenen Text, sondern auch über die Stimme des Sprechers. Dies nutzt die Sprecher-Verifikation aus, die beispielsweise für automatische Zulassungssysteme gebraucht wird, in denen die Benutzer anhand der Stimme identifiziert werden.

Grob wird zwischen der Text abhängigen und der Text unabhängigen Sprecher-Verifikation unterschieden. In der Text abhängigen Sprecher-Verifikation wird davon ausgegangen, dass in den zwei Sprachsignalen, die verglichen werden müssen, dasselbe gesprochen wurde. In diesem Fall wird oft Pattern Matching verwendet. In der Text unabhängigen Sprecher-Verifikation wird davon ausgegangen, dass in den zwei zu vergleichenden Sprachsignalen voneinander unterschiedlicher Text gesprochen wurde. Oft wird in diesem Fall mit statistischen Methoden (z.B. Gauss'schen Mischmodellen, GMMs) gearbeitet. Eine Zusammenfassung insbesondere der Text unabhängigen Sprecher-Verifikation findet sich z.B. in [1].

Meistens werden für die Sprecher-Verifikation aus dem Sprachsignal dieselben Merkmale wie für die Spracherkennung extrahiert. Es ist jedoch fraglich, ob diese Merkmale für die Sprecher-Verifikation ebenfalls optimal sind. In der Spracherkennung ist man ja gerade daran interessiert, Merkmale zu haben, die Sprecher unabhängig sind. Oder in anderen

Worten ausgedrückt, braucht man dort Merkmale, die gut zur Unterscheidung von Phonemen geeignet sind aber von der Stimmcharakteristik unabhängig sind.

Aufgabenstellung

In dieser Arbeit geht es darum, zusätzliche Merkmale wie zum Beispiel die Grundfrequenz oder die Modulation derselben aus dem Sprachsignal zu extrahieren und deren Eignung für die Sprecher-Verifikation zu evaluieren.

- Es wird empfohlen, in einem ersten Schritt Sprachmerkmale, welche mit der Grundfrequenz F_0 zusammenhängen zu untersuchen.
- Als Ausgangspunkt für die F_0-Extraktion kann ein Algorithmus, welcher am Institut entwickelt wurde, verwendet werden.
- In der Literatur soll nach geeigneten Methoden zur Extraktion von F_0-bezogenem Merkmalen wie z.B deren Modulationen gesucht werden. Als Anhaltspunkt kann z.B. [2] dienen.
- Es soll dann ein Programm-Modul zur Extraktion von F_0-bezogenen Merkmalen erstellt werden.
- Die extrahierten Merkmale sollen auf ihr Potential zur Sprecherunterscheidung hin untersucht werden. Dazu kann ein am Institut entwickeltes Rahmenprogramm verwendet werden. Um möglichst schnell eine Aussage machen zu können, wird empfohlen, die Merkmale in einer Text abhängigen Sprechererkennung basierend auf Pattern Matching zu testen. Die Wahl von geeigneten Testdaten soll mit den Betreuern abgeklärt werden.
  - In einem ersten Schritt sollen die einzelnen Merkmale separat evaluiert werden.
  - In einem zweiten Schritt sollen alle Merkmale in einem Vektor zusammengefasst werden und dieser Vektor für das Pattern Matching verwendet werden. Da die Dimensionen dieses Vektors verschiedene Einheiten haben, ist die Euklidische Distanz nicht optimal. Es kann deshalb ein Neuronales Netz trainiert werden, welches eine bessere Distanz berechnet. Siehe dazu auch [3].
- In der Fachliteratur soll nach weiteren Ansätzen für zusätzliche Sprachmerkmale für die Sprecher-Verifikation und deren Extraktion aus dem Sprachsignal gesucht werden.

Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht zu dokumentieren (siehe dazu [4]), der in gedruckter und in elektronischer Form abzugeben ist. Zusätzlich sind im Rahmen eines Kolloquiums zwei Präsentationen vorgesehen: etwa drei Wochen nach Beginn soll der Arbeitsplan und am Ende der Arbeit die Resultate vorgestellt werden. Die Termine werden später bekannt gegeben.

Literaturverzeichnis

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, and D. A. Reynolds J. Ortega-García, D. Petrovska-Delacrétaz. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, pages 430–451, April 2004.

[2] I. R. Titze. Workshop on acoustic voice analysis. National Center for Voice and Speech, 1994.

[3] M. Gerber and B. Pfister. Quasi text-independent speaker verification with neural networks. MLMI'05 Workshop, Edinburgh (United Kingdom), July 2005.

[4] B. Pfister. *Hinweise für die Präsentation der Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.

[5] B. Pfister. *Richtlinien für das Verfassen des Berichtes zu einer Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.

Zürich, den 5. April 2006 Prof. Dr. L. Thiele