Vortrag zur Studienarbeit

PLP und RASTA-PLP
Zwei Verfahren zur Vorverarbeitung von Sprachsignalen

Ekkart Bolten

am

Institut für Logik, Komplexität und Deduktionssysteme

Mai 1995

Zusammenfassung

Die beiden Verfahren *Perceptual Linear Prediction (PLP)* und *RelAtive SpecTrAl PLP (RASTA-PLP)* werden vorgestellt. Das PLP-Verfahren basiert auf der Linearen Prädiktion (LP), mit der Parameter des menschlichen Vokaltraktfilters geschätzt werden. Es erweitert die LP um Transformationen, mit denen Eigenschaften des menschlichen Hörens nachgebildet werden. Bei dem RASTA-PLP Verfahren wird zusätzliches Wissen über Eigenschaften des Kommunikationskanals genutzt, um mehr Stabilität gegenüber Variationen des Kanals zu erreichen.
Anhand eines auf Multi State Time Delayed Neural Networks (MSTDNN) basierenden Systems zur Erkennung buchstabierter Wörter werden die Verfahren auf telephonisch gesammelten Sprachdaten mit den zur Zeit am Institut verwendeten Melscale-Koeffizienten verglichen.

Inhaltsverzeichnis

1 Einführung

1.1 Sprachvorverarbeitung

Die Aufgabe maschineller Spracherkennungssysteme besteht darin, die in einem Sprachsignal kodierte Folge von Worten zu bestimmen. Diese Aufgabe wird in aktuellen Spracherkennungssystemen in zwei Teilaufgaben untergliedert:

- Sprachvorverarbeitung
  Aus dem digitalisierten Signal wird eine Folge von Merkmalvektoren extrahiert, die nach Möglichkeit nur die zur Erkennung benötigten Informationen enthalten sollen. Durch die damit erreichte Datenreduktion werden komplexere Erkennungsoperationen erst effizient realisierbar.
- Spracherkennung
  Aus der Folge von Merkmalsvektoren, die die Vorverarbeitung zu einem Sprachsignal liefert, werden Phoneme, Worte und schließlich der gesprochene Satz ermittelt.

Das Thema der Studienarbeit ist die Vorstellung und der Vergleich verschiedener Vorverarbeitungsverfahren, die sich im wesentlichen darin unterscheiden, wie sie die benötigten Informationen von redundanten trennen.
Ein Sprecher kodiert einen Satz, indem er durch die Modulation seiner Stimme Laute (Phoneme) erzeugt. Neben den Informationen über die gesprochenen Laute enthält das Signal noch redundante Informationen, die die Erkennungsleistung aktueller Erkenner stark mindern. Die Quellen dieser Redundanzen sind:

- Spektrale Eigenschaften der Sprecherstimme
  Stimmhöhe und -klangfarbe variieren bei verschiedenen Sprechern.
- Betonungen
  Schwankungen in der Stimmhöhe und Lautstärke, die sich über mehrere Phoneme erstrecken.
- Spektrale Eigenschaften des Kommunikationskanals [1]
  Statische Filtereigenschaften der Aufnahmekomponenten und dynamische Veränderungen in Abstand und Winkel zwischen Sprecher und Mikrofon.
- Störungen
  Bestandteile des Sprachsignals, die zum Beispiel durch Hintergrundgeräusche oder durch elektromagnetische Einkopplung auf dem Leitungsweg verursacht werden.

Der Mensch kann Sprache auch unter Bedingungen verstehen, bei denen aktuelle Spracherkennungssysteme überfordert sind, zum Beispiel bei lauten Hintergrundgeräuschen oder schlechten Telefonverbindungen. In der Sprachvorverarbeitung wird deshalb versucht, Eigenschaften der Erzeugung und Erkennung von Sprache durch den Menschen zu modellieren und in Verfahren zu implementieren.

1.2 Eigenschaften des menschlichen Hörens

In der digitalen Sprachvorverarbeitung werden Eigenschaften des menschlichen Hörens bei allen vorgestellten Verfahren simuliert. Dadurch sollen Signalbestandteile entsprechend ihrer Bedeutung für die Spracherkennung gewichtet werden.
Beim Hören wird das Signal durch Schädel und Gehörgang linear gefiltert und entlang der Ohrschnecke (Cochlea) durch Filterung in seine Frequenzbestandteile zerlegt. Die Energien der einzelnen Bestandteile aktivieren Rezeptorzellen an den Wänden der Cochlea zur Erzeugung von Impulsen, die über Nerven ins Gehirn geleitet werden.
Durch die (lineare) Filterung der Schallwellen bei ihrer Passage durch das Ohr und durch die nichtlineare Umwandlung von Schalldruckänderungen in Nervenimpulse durch die Rezeptorzellen resultieren folgende für die vorzustellenden Verfahren interessante Eigenschaften des menschlichen Hörens:

- Die Frequenzauflösung, also die Fähigkeit, zwei Töne ähnlicher Frequenz unterscheiden zu können, nimmt nichtlinear bei steigenden Frequenzen ab.
- Die Lautstärkenauflösung nimmt bei steigender Signalenergie nichtlinear ab.
- Gegenüber Tönen steigender Frequenz nimmt die Empfindlichkeit des Gehörs oberhalb einer Grenzfrequenz nichtlinear ab.

[1] Kommunikationskanal: Analoge Komponenten vom Mikrofon bis zum AD-Wandler (siehe auch 1.4 auf Seite 5)

1.3 Synthese von Sprache durch ein lineares System

In diesem Modell wird das spracherzeugende System untergliedert in die Erzeugung eines Anregungssignals und die darauf folgende Modulation des Signals durch lineare Operationen. Beide Bestandteile werden im Sprachsignal in diesem Modell zeitabhängig moduliert. Das Blockschaltbild der diskreten Version[2] des Systems ist der Abbildung 1 zu entnehmen.

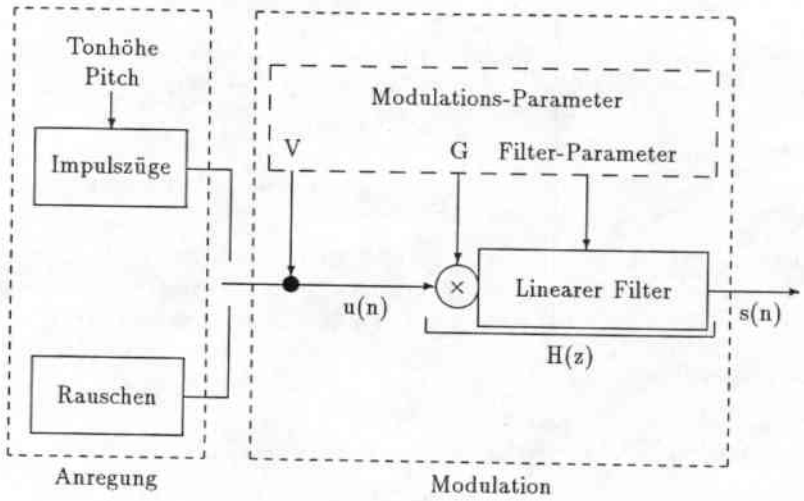


Abbildung 1: Diskretes Spracherzeugungsmodell

Es können zwei verschiedene Anregungssignale durch den Menschen erzeugt werden:

- Rauschen für stimmlose Laute
  In starken Verengungen des Vokaltraktes wird durch Luftverwirbelungen ein Rauschen erzeugt. Im Leistungsspektrum (siehe 2.4 auf Seite 7) sind die Energien annähernd gleichverteilt.
- Impulszüge für stimmhafte Laute
  Durch Luftströmungen werden die Stimmbänder zu harmonischen Schwingungen angeregt. Betrachtet man das Leistungsspektrum des so erzeugten Signals, so sind Spitzen an den Frequenzen des Grundtons (Pitch) und der Obertöne der Stimmbandschwingungen zu erkennen. Man spricht deshalb von einem Impulszug-Signal.

Im Schaltbild wird die Auswahl zwischen den Anregungssignal durch die Stellung eines Schalters angegeben. Je nachdem, ob der Impulszug- oder der Rauschgenerator das Signal u(n) liefert, wird die Schalterstellung V auf voiced (stimmhaft) oder unvoiced (stimmlos) gesetzt.
Durch den Druck, mit dem Luft durch Verengungen oder Stimmbänder gepreßt wird, kann die Intensität des Signals variiert werden, was im Blockschaltbild durch Aufmultiplikation eines Verstärkungsfaktors G modelliert wird.
Das so erzeugte Anregungssignal passiert anschließend den Vokaltrakt, in dem es linear gefiltert wird. Durch eine Veränderung der Stellung von Unterkiefer, Zunge und Lippen kann diese Filterung zeitlich variiert werden. Im Modell wird das Anregungssignal durch einen (digitalen) linearen Filter gefiltert. Zusammen mit der Verstärkung um den Faktor G kann die Filterung im Vokaltrakt beschrieben werden durch die folgende Transferfunktion:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_k z^{-i}} \tag{1}$$

Verstärkung und Filterung können berechnet werden durch:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{2}$$

Die Modellierung durch einen linearen Filter, der nur Pole und keine Nullstellen besitzt, beschreibt den menschlichen Vokaltraktfilter vor allem bei der Bildung von stimmhaften und stimmlosen Lauten. Nasal- und Plosivlaute werden durch

[2] In der mathematische Behandlung des Modelles werden diskrete Signale und Funktionen benutzt. $u(k)$ und $s(k)$ für $k = 0, 1, 2, 3...$ entspre im Analogen den Zeitfunktionen $u_a(t)$ und $s_a(t)$, für $t \geq 0$.

Filter, die zusätzlich Nullstellen besitzen, besser modelliert. Ist die Ordnung des Systems jedoch groß genug, so kann der hier angegebene sogenannte All-Pole-Filter auch diese Laute hinreichend gut beschreiben.

Die zeitliche Modulation des eigentlichen Sprachsignals wird in dem Modell zu einem bestimmten Zeitpunkt durch die Werte der Modulations-Parameter V, G und $(a_i)_{i=1}^{p}$ charakterisiert. Diese Parameter können aus dem Sprachsignal mit der Linearen Prädiktion (Abschnitt 4.1, Seite 9) geschätzt werden.

1.4 Der Kommunikationskanal

Der Kommunikationskanal besteht aus den analogen Komponenten, in denen das Signal durch Störungen oder spektrale Verzerrung (Filterung) verändert werden kann:

- der Raum, in dem sich Sprecher und Mikrofon befinden
- das Mikrofon
- die elektrische Leitung zwischen Mikrofon und AD-Wandler
- der AD-Wandler

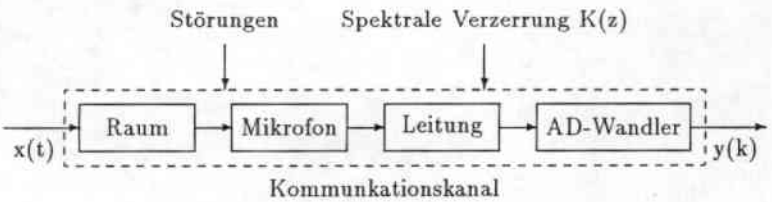


Abbildung 2: Kommunikationskanal

Störungen zeichnen sich dadurch aus, daß dem Sprachsignal ein nicht mit diesem korreliertes Signal additiv aufgeprägt wird. Sie können zum Beispiel durch Hintergrundgeräusche im Aufnahmeraum, das Übersprechen von anderen Signalen auf dem Leitungsweg oder durch elektromagnetische Störimpulse verursacht werden. Im allgemeinen lassen sich diese Störungen nur schwer modellieren und werden deshalb auch bei den hier besprochenen Verfahren nicht berücksichtigt.

Spektrale Verzerrungen durch den Kommunikationskanal entsprechen einer Filterung des analogen Sprachsignals mit der Transferfunktion $K(z)$. Zusammen mit der Transferfunktion $H(z)$ aus dem Synthesemodell ergibt sich die Transferfunktion:

$$\dot{H}(z) = H(z)\ K(z)$$

$K(z)$ kann variieren, zum Beispiel wenn unterschiedliche elektrische Leitungen benutzt werden oder wenn sich die Position des Sprechers im Vergleich zum Mikrofon ändert. Im allgemeinen sind diese Variationen im Vergleich zu denen der modulierten Filterung des Signals im Vokaltrakt so langsam, daß sie als annähernd konstant angesehen werden können.
Es gibt unterschiedliche Ansätze, diese langsamer variierende Filterung durch den Kommunikationskanal in der Sprachvorverarbeitung abzuschätzen und ihren Einfluß auf die extrahierten Merkmale zu kompensieren. Diese Kompensation ist vor allem dann notwendig, wenn Spracherkennungssysteme mit wechselnden Kommunikationskanälen betrieben werden sollen, wie zum Beispiel bei automatischen telefonischen Auskunftsdiensten.
Werden alle Sprachdaten für ein Erkennungssystem über die gleichen analogen Komponenten aufgenommen und digitalisiert, haben die spektralen Eigenschaften des Kanals nur geringe Wirkung auf die Erkennungsleistung, so daß eine Kompensation in aktuellen Spracherkennern, die auf neuronalen Netzen oder LVQ's basieren, entfallen kann.

2 Begriffe

2.1 Abtastung

Ausgangspunkt für die Erkennung eines gesprochenen Satzes ist das analoge Sprachsignal $s_a(t)$ in einem Zeitintervall $[0, t_{max}]$. Zur digitalen Verarbeitung wird eine diskrete Folge von Amplitudenwerten $s(n)$ durch Abtastung des analogen Sprachsignals mit der Abtastrate ΔT (Abtastfrequenz $f_s = \frac{1}{\Delta T}$) ermittelt:

$$s(n) = s_a(n\ \Delta T) \qquad 0 \leq n < n_{\max}$$

mit

$$n_{\max} = \left\lceil \frac{t_{\max}}{\Delta T} \right\rceil$$

Nach dem Abtasttheorem muß die Abtastfrequenz mindestens das Doppelte der maximalen Frequenz im Signal betragen, um eine Rekonstruktion des Signals ohne Informationsverluste zu ermöglichen. Übliche Abtastfrequenzen reichen von 8kHz bis 20kHz, so daß maximale Frequenzen im Sprachsignal von 4kHz bis 10kHz erfaßt werden.

2.2 Kurzzeitanalysen

Die für die Spracherkennung benötigten Informationen sind nach dem Modell zur Sprachsynthese in der zeitlichen Variation der Modulationsparameter kodiert. Da die Variationen durch mechanische Veränderungen im Vokaltrakt verursacht werden, können sie nicht beliebig schnell erfolgen. Die maximale Variationsfrequenz liegt bei ca. 100Hz. Analysiert man das Sprachsignal in hinreichend kleinen Zeitintervallen, so kann man dort von konstanten Modulationsparametern ausgehen und diese direkt oder indirekt durch ihre Wirkung auf die spektrale Zusammensetzung des Signalabschnittes bestimmen. Dazu wird das Sprachsignal in meist überlappende, äquidistant verteilte Abschnitte unterteilt, die im folgenden Analysefenster genannt werden. Ein Analysefenster beinhaltet N Werte und wird durch Multiplikation des Sprachsignals mit einer Fensterfunktion berechnet.
Übliche Fensterfunktionen sind das Rechteck-Fenster $w_R(n)$ und das sogenannte Hamming-Fenster $w_H(n)$:

$$w_R(n) = \begin{cases} 1 & 0 \leq n < N \\ 0 & \text{sonst} \end{cases}$$

$$w_H(n) = \begin{cases} 0.54 - 0.46cos\left(2\pi \frac{n}{N}\right) & 0 \leq n < N \\ 0 & \text{sonst} \end{cases}$$

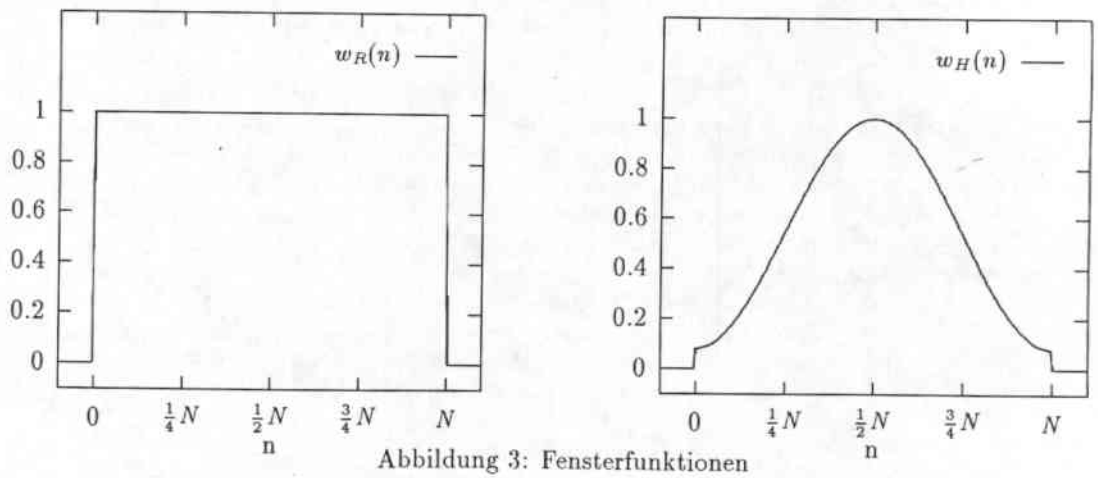


Abbildung 3: Fensterfunktionen

Das Hamming-Fenster wird wegen seiner besseren spektralen Eigenschaften bei allen vorzustellenden Verfahren benutzt. Bei einem Abstand von Δn Abtastwerten zwischen zwei Analysefenstern gilt für das Analysefenster $s_m(n)$:

$$s_m(n) = s(m\ \Delta n + n)\ w(n) \qquad 0 \leq n < N \qquad 0 \leq m \leq \left\lfloor \frac{n_{max} - N}{\Delta n} \right\rfloor \tag{3}$$

2.3 Das Kurzzeit-Frequenzspektrum

Um eine Frequenzanalyse ähnlich der im menschlichen Gehör zu realisieren, wird das Signal aus dem Zeitbereich in den Frequenzbereich transformiert. Dazu werden aus einem Sprachfenster $s_m(n), 0 \leq n < N$ die komplexen Koeffizienten $S_m(k), 0 \leq k < N$ des sogenannten Kurzzeit-Frequenzspektrums errechnet.
Die Gleichungen der Diskreten-Fourier-Transformation (DFT)

$$\begin{aligned} S_m(k) &= \sum_{n=0}^{N-1} s_m(n)\, e^{-j\frac{2\pi}{N}kn} \\ s_m(n) &= \frac{1}{N}\sum_{k=0}^{N-1} S_m(k)\, e^{j\frac{2\pi}{N}kn} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} S_m(k)\, \left(\cos(\frac{2\pi}{N}kn) + j\, \sin(\frac{2\pi}{N}kn)\right) \end{aligned} \tag{4}$$

geben den Zusammenhang zwischen dem Spektrum und dem Sprachfenster an, wenn die Bedingungen des Abtasttheorems erfüllt sind. In der Implementierung der Frequenzanalyse werden die Koeffizienten $S_m(k)$ oder äquivalente mit der Fast-Fourier-Transformation (FFT) oder der effizienteren Fast-Hartley-Transformation (FHT) berechnet.
Aus Gleichung (4) ist zu ersehen, daß die Transformation das ursprüngliche Signal als Summe von Sinus- und Cosinusschwingungen darstellbar macht. Die komplexen Koeffizienten $S_m(k)$ beschreiben dabei den Anteil, den die Schwingungen bei der Frequenz f_k haben (f_s ist dabei wieder die Abtastfrequenz):

$$f_k = \frac{k}{N}\frac{f_s}{2} \tag{5}$$

Das ausgeschnittene und gewichtete Signalfenster ist eine Folge reeller Zahlen. Bei reellen Eingaben liefert die Fouriertransformation ein Spektrum mit folgenden Symmetriebedingungen für $1 \leq k < \frac{N}{2}$:

$$\begin{aligned} Re\,(S_m(k)) &= Re\,(S_m(N-k)) \\ Im\,(S_m(k)) &= -Im\,(S_m(N-k)) \end{aligned}$$

2.4 Das Kurzzeit-Leistungsspektrum

Die Koeffizienten des Kurzzeit-Leistungsspektrums $P_m(k)$ werden aus den Koeffizienten $S_m(k)$ des Kurzzeit-Frequenzspektrums errechnet:

$$P_m(k) = Im\,(S_m(k))^2 + Re\,(S_m(k))^2$$

Für die Koeffizienten $P_m(k)$ gilt die Symmetrie $P_m(k) = P_m(N-k)$ für $1 \leq k < \frac{N}{2}$.
Wie im vorigen Abschnitt zu sehen war, kann das Sprachsignal aus den Koeffizienten des Frequenzspektrums $S_m(k)$ durch Kombination von harmonischen Schwingungen rekonstruiert werden. Mit der Berechnung des Leistungsspektrums wird die Rekonstruierbarkeit aufgegeben. Die Koeffizienten $P_m(k)$ beschreiben den gemeinsamen Anteil der Cosinus- und Sinusschwingung der Frequenz f_k (Gleichung (5)) am Gesamtsignal. Die durch diese Transformation verlorenen Informationen beschreiben im Wesentlichen die Phaseneigenschaften des Vokaltraktfilters. Informationen über die Dämpfung im Vokaltrakt bleiben erhalten und sind für die weitere Spracherkennung gut geeignet.

3 Melscale-Analyse

Zur Zeit werden im Institut aus den abgetasteten Sprachdaten Melscale-Koeffizienten ermittelt, deren Berechnung im folgenden kurz vorgestellt wird:
Zunächst werden aus einem Sprachfenster $s_m(n)$ der Breite N die Koeffizienten des Kurzzeit-Leistungsspektrums $P_m(k)$, $0 \leq i < \frac{N}{2}$ errechnet (siehe 2.4).
Durch Zusammenfassung der $P_m(k)$ über N_{Int} Intervalle (Bandbreiten) $[u_i, o_i]$ mit $0 \leq u_i \leq o_i < \frac{N}{2} - 1$, $0 \leq i < N_{Int}$ werden die Melscale-Koeffizienten $M_m(i)$ ermittelt:

$$M_m(i) = \log\left(\frac{1}{2}P_m(u_i) + \left(\sum_{k=u_i+1}^{o_i-1} P_m(n)\right) + \frac{1}{2}P_m(o_i)\right) \qquad 0 \leq i < N_{Int}$$

Dabei überlappen sich die Intervalle jeweils an den Grenzen: $o_i = u_i$ für $1 \leq i < N_{Int}$.
Mit den zur Zeit verwendeten Parametern

- Abtastfrequenz $f_s = 16kHz$
- Fensterbreite $N = 256$
- Intervallanzahl $N_{Int} = 16$

ergibt sich die der Abbildung 4 zu entnehmende Verteilung der Koeffizienten auf die verschiedenen Frequenzbereiche.

Koeffizient $M_m(i)$ i	Intervallanfang Hz	Intervallende Hz
0	0	125
1	125	375
2	375	625
3	625	875
4	875	1125
5	1125	1375
6	1375	1625
7	1625	1875
8	1875	2187.5
9	2187.5	2563.5
10	2563.5	3000
11	3000	3562.5
12	3562.5	4250
13	4250	5062.5
14	5062.5	6062.5
15	6062.5	7250

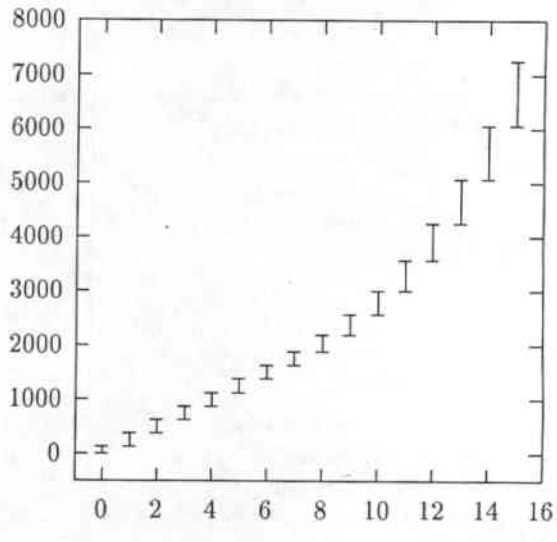


Abbildung 4: Verteilung der Melscale-Koeffizienten

Bei der Berechnung von Melscale-Koeffizienten werden die logarithmische Lautstärkeempfindlichkeit und die Frequenzauflösung des menschlichen Gehörs nachgebildet. Die frequenzabhängige Lautstärkeempfindlichkeit wird nicht berücksichtigt.

4 Verfahren zur Schätzung der Vokaltraktparameter

Die im folgenden beschriebenen Verfahren basieren auf dem Linearen Sprachsynthesemodell (siehe 1.3). Im Gegensatz zur Melscale-Analyse, deren Koeffizienten eine Aussage über die spektrale Zusammensetzung des Signals darstellen, beschreiben die berechneten Koeffizienten die Ursache dieser Zusammensetzung, nämlich die Filterfunktion des Vokaltraktes.

4.1 Linear Prediction (LP)

Bei der LP (oft auch als Linear Predictive Coding (LPC) bezeichnet) werden für ein Analysefenster $s_m(n)$ zum Zeitpunkt m die Koeffizienten $\alpha_{m,i}$ eines linearen Systems, des sogenannten Linearen Prädiktors, so ermittelt, daß es das Sprachsignal im Fenster nach einem noch anzugebenden Kriterium optimal vorhersagt.
Für ein Kurzzeit-Analysefenster (siehe 2.2 (3)) $s_m(n)$ der Breite N ist ein linearer Prädiktor der Ordnung p mit Koeffizienten $\alpha_{m,i}$ gegeben durch die Gleichung

$$\tilde{s}_m(n) = \sum_{i=1}^{p} \alpha_{m,i}\ s_m(n-i) \tag{6}$$

Der Schätzfehler $e_m(n)$ und seine Z-Transformierte $E_m(z)$ lauten:

$$e_m(n) \;=\; s_m(n) - \tilde{s}_m(n) \;=\; s_m(n) - \sum_{i=1}^{p} \alpha_{m,i}\ s_m(n-i) \tag{7}$$

$$E_m(z) \;=\; S_m(z)\ A(z) \qquad \text{mit}\ \ A(z) = 1 - \sum_{i=1}^{p} \alpha_{m,i}\ z^{-i} \tag{8}$$

$A(z)$ ist die Transferfunktion eines linearen Filters, der bei Eingabe des Sprachsignals den Schätzfehler ermittelt.
Die Koeffizienten $\alpha_{m,i}$ werden durch Lösung[3] eines linearen Gleichungssystems so bestimmt, daß der quadratische Fehler EM_m über dem Bereich I=[0...N-1] minimiert wird.

$$EM_m = \sum_{n=0}^{N-1} e_m(n)^2 = \sum_{n=0}^{N-1} (s_m(n) - \tilde{s}_m(n))^2 \tag{9}$$

Man kann die so bestimmten Koeffizienten $\alpha_{m,i}$ als Schätzung für die Filterkoeffizienten a_i des linearen Filters im Sprachsynthesemodell (siehe 1.3 (1) und (2) auf Seite 4) betrachten. Dann kann das Fehlerfiltersystem mit Transferfunktion $A(z)$ als inverser Filter zum Vokaltraktfilter interpretiert werden, da

$$A(z) = \frac{G}{H(z)}$$

Hat man die $\alpha_{m,i}$ berechnet, so können auch die anderen Vokaltraktparameter relativ einfach bestimmt werden. Betrachtet man den Schätzfehler des Linearen Prädiktors mithilfe der Gleichung (2) Seite 4) wie folgt genauer

$$e_m(n) = s_m(n) - \tilde{s}_m(n) = \sum_{i=1}^{p} a_i s_m(n-i) + Gu(n) - \sum_{i=1}^{p} \alpha_{m,i}\ s_m(n-i)$$

so ergibt sich bei identischen Koeffizienten ($a_i = \alpha_{m,i}$):

$$e_m(n) = G\ u(n)$$

Aus dieser Beziehung lassen sich G und die Anregungsart (Rausch- oder Impulszug-Anregung) bestimmen.

Setzt man die geschätzten Koeffizienten $\alpha_{m,i}$ und G in die Gleichung des linearen Filters im Sprachsynthesemodell ein

$$H_m(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_{m_k} z^{-k}}$$

[3] In [6] werden drei Lösungsverfahren zur Berechnung der Koeffizienten des Linearen Prädiktors angegeben. In den im folgenden beschriebene Verfahren wird die Autokorrelations-Methode nach Durbin verwendet.

und bestimmt das Verstärkungsverhalten des Filters durch Berechnung von

$$|H_m(z)| \text{ mit } z = e^{j2\pi f},\ 0 \leq f < \frac{f_s}{2}$$

für die Frequenzen bis zur halben Abtastfrequenz f_s, so erhält man wie beim Kurzzeit-Leistungsspektrum wieder eine Darstellung, in der die Energieverteilung im Sprachsignal angegeben ist. Man kann die Lineare Prädiktion als Approximation des Kurzzeit-Leistungsspektrums interpretieren. Im Vergleich zu diesem, ist das Verstärkungsverhalten je nach Modellordnung p stark geglättet. Feinere harmonische Strukturen, die von Sprecher zu Sprecher variieren, werden so unterdrückt, während die für die Erkennung notwendigen Resonanzen im Vokaltrakt (Formanten) gut dargestellt werden. Die Modellordnung p des Linearen Prädiktors bestimmt die Genauigkeit der Approximation. Je nach Aufgabe (zum Beispiel sprecherabhängige oder sprecherunabhängige Erkennung) muß eine geeignete Ordnung bestimmt werden, bei der das Verfahren durch Glättung möglichst viel redundante Informationen unterdrückt, aber die zur Erkennung benötigten Informationen über Resonanzen im Vokaltraktfilter bewahrt.[4]

Im allgemeinen werden die mit der LP geschätzten Koeffizienten nicht direkt zur Erkennung benutzt, sondern in andere Darstellungen transformiert. Bei der PLP und der RASTA-PLP werden aus den Koeffizienten Cepstral-Koeffizienten berechnet.

4.2 Perceptual Linear Prediction (PLP)

Bei der PLP wird das Sprachsignal zunächst transformiert, um Eigenschaften des menschlichen Gehörs nachzubilden, das verschiedene Frequenz- und Lautstärkebereiche unterschiedlich genau auflöst. Aus dem transformierten Signal werden wie bei der LP aus den Koeffizienten eines linearen Prädiktors die Cepstral-Koeffizienten ermittelt.

Frequenzanalyse. Aus dem Sprachsignal $s_a(t)$ [5] wird mit dem analogen Hamming-Fenster $w_{Ha}(t)$ ein Analysefenster s_{t_0} der Breite T ausgeschnitten:

$$\begin{aligned} w_{Ha}(t) &= \begin{cases} 0.54 - 0.46\cos 2\pi t & 0 \leq t < 1 \\ 0 & \text{sonst} \end{cases} \\ s_{t_0}(t) &= s_a(t_0 + t)\ w_{Ha}\left(\frac{t}{T}\right) \qquad 0 \leq t \leq T \end{aligned} \tag{10}$$

Die Fourier-Transformierte S_{t_0} und das analoge Leistungsspektrum P_{t_0} des Analysefensters sind wie folgt definiert:

$$\begin{aligned} S_{t_0}(\omega) &= \mathcal{F}\{s_{t_0}(t)\}(\omega) = \int_0^T s_{t_0}(t)\ e^{-jt\omega} dt \\ P_{t_0}(\omega) &= Im(S_{t_0}(\omega))^2 + Re(S_{t_0}(\omega))^2 \end{aligned}$$

Das Leistungsspektrum $P_{t_0}(\omega)$ beschreibt die Energieverteilung im Signal abhängig von Kreisfrequenzen $\omega = 2\pi f$.

Transformation in die Bark-Scale. Die Auflösung des menschlichen Gehörs in der Kreisfrequenz-Skala nimmt mit steigenden Frequenzen ab. Bei der PLP wird das Leistungsspektrum durch eine bei steigenden Frequenzen stärker werdende Stauchung entlang der Frequenzachse in die Bark-Scale transformiert

$$\dot{P}_{t_0}(\Omega) = P_{t_0}(T_\omega(\Omega))$$

mit den Transformationsfunktionen (siehe Abbildung 5)

$$T_\Omega(\omega) = 6\ \log\left(\frac{\omega}{1200\ \pi} + \sqrt{1 + \left(\frac{\omega}{1200\ \pi}\right)^2}\right) \quad \text{und} \quad T_\omega(\Omega) = 1200\ \pi \sinh(\frac{\Omega}{6}) \tag{11}$$

In der Bark-Scale ist die Frequenzauflösung des menschlichen Hörens über alle Bark-Frequenzen gleich. Die Abtastfrequenz in der Bark-Scale ist

$$b_s = T_\Omega(2\pi f_s)$$

[4] In [2] wird die Ordnung $p = 14$ als optimal für eine Erkennungsaufgabe angegeben, bei der ein Erkenner auf einen Sprecher trainiert und einem anderen getestet wurde.

[5] Zur besseren Darstellung der Zusammenhänge wird die Berechnung der PLP -Koeffizienten zunächst im Analogen und danach in der diskr Implementierung angegeben.

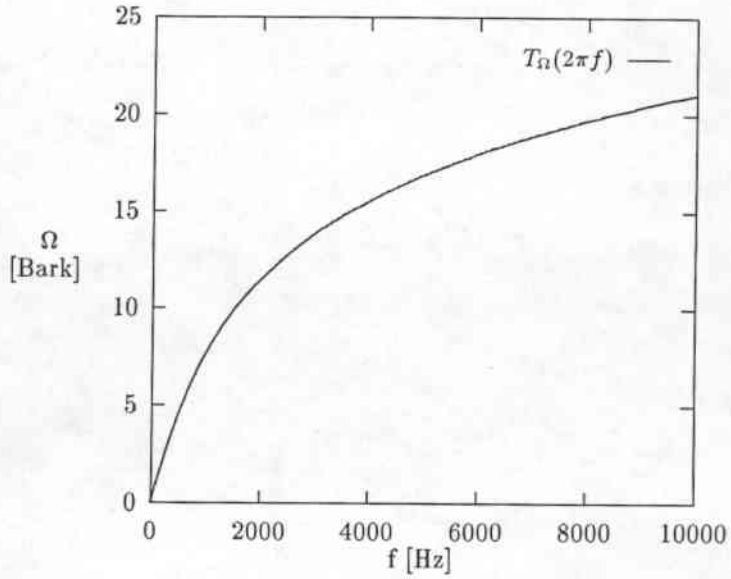


Abbildung 5: Bark-Transformationsfunktion

Berechnung des Auditiven Spektrums. Das transformierte Leistungsspektrum wird mit der Kritischen-Band-Funktion $\Psi(\Omega)$ (siehe Abbildung 6) gefaltet. Das Ergebnis der Faltung ist das Auditive Spektrum des Sprachsignals:

$$\Theta_{t_0}(\Omega) = \int_{-1.3}^{2.5} \dot{P}_{t_0}(\tilde{\Omega} - \Omega)\ \Psi(\tilde{\Omega})\ d\tilde{\Omega} \tag{12}$$

$$\Psi(\Omega) = \begin{cases} 10^{2.5(\Omega+0.5)} & -1.3 \leq \Omega \leq -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{sonst} \end{cases} \tag{13}$$

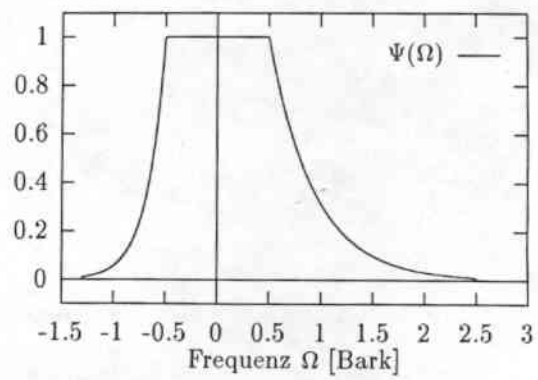


Abbildung 6: Kritische-Band-Funktion

Auswertung des Auditiven Spektrums an einzelnen Frequenzen. Durch die Faltung mit der relativ breiten Kritischen-Band-Funktion wird die spektrale Auflösung des Auditiven Spektrums im Vergleich zum Leistungsspektrum stark reduziert. Es reicht, $\Theta_{t_0}(\Omega)$ an einzelnen Bark-Frequenzen Ω_i auszuwerten. Bei der PLP wird $\Theta_{t_0}(\Omega)$ über den Frequenzbereich von 0 Bark bis zur halben Abtastfrequenz b_s zB. 18 Bark (bei 16kHz Abtastfrequenz) in Abständen von 0.994 Bark ermittelt und beschreibt damit das Signal durch die Folge $(\Theta_{t_0,i} = \Theta_{t_0}(\Omega_i))$, deren Zusammenhang zu den

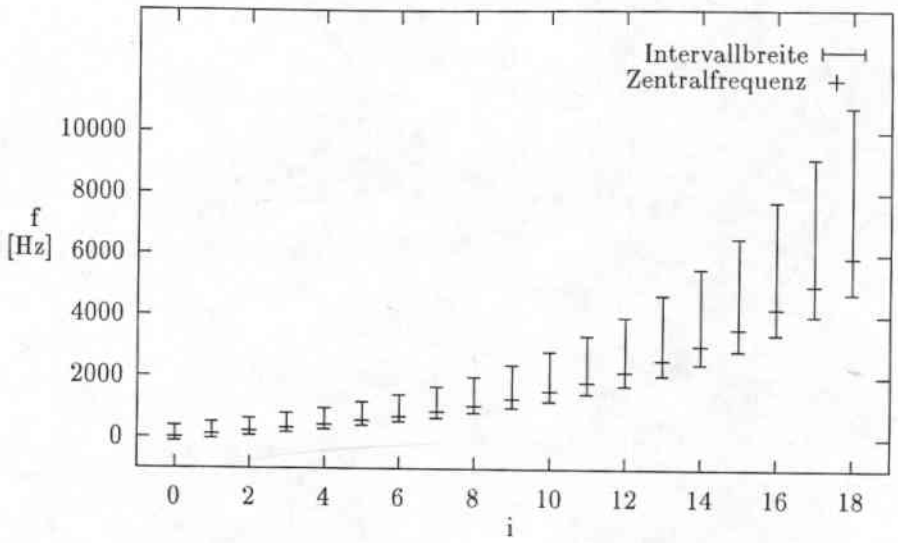


Abbildung 7: Verteilung der $\Theta_{t_0,i}$

überdeckten Intervallen in der Abbildung 7 gezeigt wird.

$$(\Theta_{t_0,i} = \Theta_{t_0}(\Omega_i))_{i=0}^{i_{\max}} \quad \text{mit} \quad \Omega_i = 0.994i, \; i_{\max} = \left\lfloor \frac{b_s}{0.994} \right\rfloor$$

Anpassung an die frequenzabhängige Lautstärkeempfindlichkeit des Ohres. Um die in verschiedenen Frequenzbereichen unterschiedliche Empfindlichkeit des menschlichen Gehörs zu simulieren, werden die $\Theta_{t_0,i}$ mit einer ebenfalls in die Bark-Scale transformierten Equal-Loudness-Curve $E(\omega)$ multipliziert:

$$\Xi_{t_0,i} = \Theta_{t_0,i} \; E\left(T_\omega\left(\Omega_i\right)\right)$$

wobei je nach benutzter Abtastfrequenz f_s eine der Kurven E_1 oder E_2 (siehe Abbildung 8) benutzt wird:

$$E_1(\omega) = \frac{(\omega^2 + 56.8\ 10^6)\ \omega^4}{(\omega^2 + 6.3\ 10^6)^2\ (\omega^2 + 0.38\ 10^9)} \qquad f_s \leq 10\text{kHz} \tag{14}$$

$$E_2(\omega) = \frac{(\omega^2 + 56.8\ 10^6)\ \omega^4}{(\omega^2 + 6.3\ 10^6)^2\ (\omega^2 + 0.38\ 10^9)\ (\omega^6 + 9.58\ 10^{26})} \qquad f_s > 10\text{kHz} \tag{15}$$

Anpassung an die Lautstärkeauflösung des Ohres. Die nichtlineare Empfindlichkeit des menschlichen Gehörs gegenüber Tönen unterschiedlicher Lautstärke wird durch das Ziehen der 3. Wurzel aus den $\Xi_{t_0,i}$ kompensiert, was in [2] als Anwendung des *Power Law of Hearing* bezeichnet wird.

$$\Phi_{t_0,i} = \sqrt[3]{\Xi_{t_0,i}} \qquad i = 0, 1, 2..., i_{\max}$$

Die Koeffizienten $\Phi_{t_0,i}$ werden mit der inversen DFT (IDFT) in den Zeitbereich zurücktransformiert. Das Ergebnis ist ein auf wenige Amplitudenwerte reduziertes Zeitsignal, aus dem mithilfe der Linearen Prädiktion und einer cepstralen Rekursion die PLP-Koeffizienten berechnet werden.
Durch die Informationsverdichtung der PLP bei der Transformation des Sprachsignals werden niedrigere Modellordnungen p des Prädiktors benötigt, um das Signal zu beschreiben, als bei der LP. So wird zum Beispiel in [2] $p = 8$ als optimale Ordnung bei der PLP gegenüber $p = 14$ bei der LP angegeben, um ein auf einem Sprecher trainierten Spracherkenner auf einen anderen Sprecher anzuwenden.

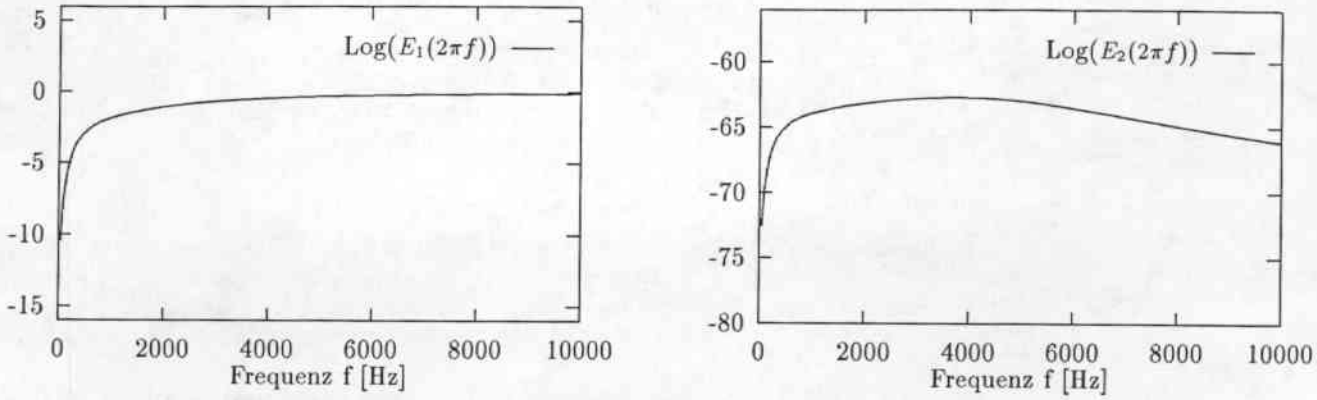


Abbildung 8: Equal-Loudness Kurven

Implementierung. Die im Analogen dargestellte Berechnung des Auditiven Spektrums durch das Faltungsintegral (Gleichung 12) muß in der Implementierung auf einem Rechner ersetzt werden durch die Berechnung von gewichteten Summen über die Koeffizienten des diskreten Kurzzeit-Leistungsspektrums $P_m(k)$, die das Signal jeweils bei der Frequenz f_k in Hertz bzw. b_k in Bark beschreiben:

$$f_k = \frac{k}{N}\frac{f_s}{2} \qquad \text{und} \qquad b_k = T_\Omega(f_k)$$

Für die Berechnung von $\Theta_m(\Omega)$ werden Koeffizienten $P_m(k)$ benutzt, deren Indizes in der Menge $M(\Omega)$ enthalten sind:

$$M(\Omega) = \{\ k \mid -1.3 \leq b_k - \Omega \leq 2.5\}$$

Dann ergeben sich die diskreten Berechnungen von $\Theta_m(\Omega)$ und die der Koeffizienten $\Phi_{m,i}$ durch:

$$\Theta_m(\Omega) = \sum_{k \in M(\Omega)} P_m(k)\Psi(\Omega - b_k)$$

$$\Phi_{m,i} = \sqrt[3]{\Theta_m(\Omega_i)\ E(T_\omega(\Omega_i))} \qquad i = 0, 1, 2..., i_{\max}$$

Dabei werden die Gewichte $\Psi(\Omega_i - b_k)$ und die Werte $E(T_\omega(\Omega_i))$ in der Implementierung für alle $k \in M(\Omega_i)$ und $i = 0, 1, 2..., i_{\max}$ im Vorraus berechnet und zur Berechnung der Koeffizienten $\Phi_{m,i}$ in jedem Analysefenster m benutzt.

4.3 RelAtive SpecTrAl Perceptual Linear Prediction (RASTA-PLP)

Im allgemeinen sinkt die Erkennungsleistung von Spracherkennungs-Systemen bei der Verwendung unterschiedlicher Kommunikationskanäle stark ab. Durch die unterschiedliche Filterung des Sprachsignals in den Kanälen können Erkennungsmerkmale im Signal variieren. Beim RASTA-PLP -Verfahren wird versucht, die Auswirkungen der spektralen Verzerrung des Signals im Kommunikationskanal zu reduzieren.

4.3.1 Berechnung von RASTA-PLP-Koeffizienten

Im Abschnitt 1.4 (Seite 5) wurde ein Modell des Kommunikationskanals erläutert. Dabei wurde die Verzerrung im Kanal als lineare Filterung des Signals mit einer im Vergleich zum Vokaltrakt langsamer variierenden Transferfunktion dargestellt. Um Sprachmodulation und Kanaleinfluß zu trennen, wird die PLP um einen Zwischenschritt erweitert, in dem konstante und langsam variierende spektrale Merkmale im Signal reduziert werden.

Zusammenhang zwischen Vokaltrakt- und Kanalfilter. Im Kommunikationskanal wird das Sprachsignal mit der Übertragungsfunktion des Kanalfilters gefaltet:

$$\begin{aligned} \dot{s}(t) &= \int_{-\infty}^{\infty} s(t) f(n-t) dt \qquad \text{analog} \\ \dot{s}(n) &= \sum_{k=-\infty}^{\infty} s(k) f(n-k) \qquad \text{diskret} \end{aligned}$$

Durch die Eigenschaft der Fourier-Transformation, Faltungen im Zeitbereich auf Multiplikationen im Frequenzbereich abzubilden, ergeben sich im diskreten Frequenzbereich und im diskreten, logarithmierten Frequenzbereich die folgenden Zusammenhänge zwischen dem Kurzzeitspektrum S_m des Vokaltraktsignals, dem Kurzzeitspektrum $\dot{S}_m$ des gefilterten Signals und der Transferfunktion F_m des Kanals

$$\begin{aligned} \dot{S}_m(k) &= S_m(k)\ F_m(k) \\ \log\left|\dot{S}_m(k)\right| &= \log|S_m(k)| + \log|F_m(k)| \end{aligned}$$

Im logarithmierten Frequenzbereich sind Modulation des Sprachsignals durch den Vokaltrakt und Verzerrung durch den Kommunikationskanal additive Komponenten unterschiedlicher Variationsfrequenzen.

Trennung der additiven Komponenten. Die Idee bei der RASTA-PLP besteht darin, das logarithmierte Auditive Spektrum[6] $\Theta_{m,i} = \Theta_m(\Omega_i)$ der PLP zunächst durch eine Regressionsgerade über 5 zeitlich aufeinanderfolgende Spektren zu differenzieren

$$y_{m,i} = 0.1\,(2\log|\Theta_{m,i}| + \log|\Theta_{m,i-1}| - \log|\Theta_{m,i-3}| - 2\log|\Theta_{m,i-4}|)$$

und anschließend wieder zu integrieren

$$\dot{y}_{m,i} = y_{m,i} + 0.98\ \dot{y}_{m,i-1}$$

Die Werte $\dot{y}_{m,i}$ werden exponentiert

$$\dot{\Theta}_{m,i} = e^{\dot{y}_{m,i}}$$

Das Ergebnis ist ein in seinen konstanten und langsam variierenden Bestandteilen reduzierte Auditives Spektrum, das wie bei der PLP dazu benutzt wird, die Koeffizienten eines linearen Prädiktors zu bestimmen.

Mathematisch entspricht die Kombination von Differentiation und Integration dem Eliminieren konstanter Bestandteile. Die diskrete Implementierung der beiden Operationen realisiert jedoch eine Bandpaßfilterung, deren Eigenschaften im nächsten Abschnitt genauer angegeben werden.

[6] Die Koeffizienten $\Theta_m(\Omega_i)$ werden durch gewichtete Addition mehrerer Kurzzeitspektrums-Koeffizienten des digitalisierten Sprachsigna mittelt. Durch diese Berechnung wird aber der rein multiplikative beziehungsweise logarithmiert, additive Zusammenhang zwischen Vokalt und Kanalfilterung verloren. Ein Informationsverlust im Vergleich zur Filterung des Leistungsspektrums ist zu erwarten.

4.3.2 Eigenschaften des Bandpaßfilters

Beide numerische Operationen, Differentiation und Integration, entsprechen Filterungen des logarithmierten Auditiven Spektrums mit den Transferfunktionen

$$H_{\text{Diff.}}(z) = 0.1\left(2 + z^{-1} - z^{-3} - 2z^{-4}\right) \quad \text{und} \quad H_{\text{Int.}}(z) = \frac{1}{1 - 0.98z^{-1}}$$

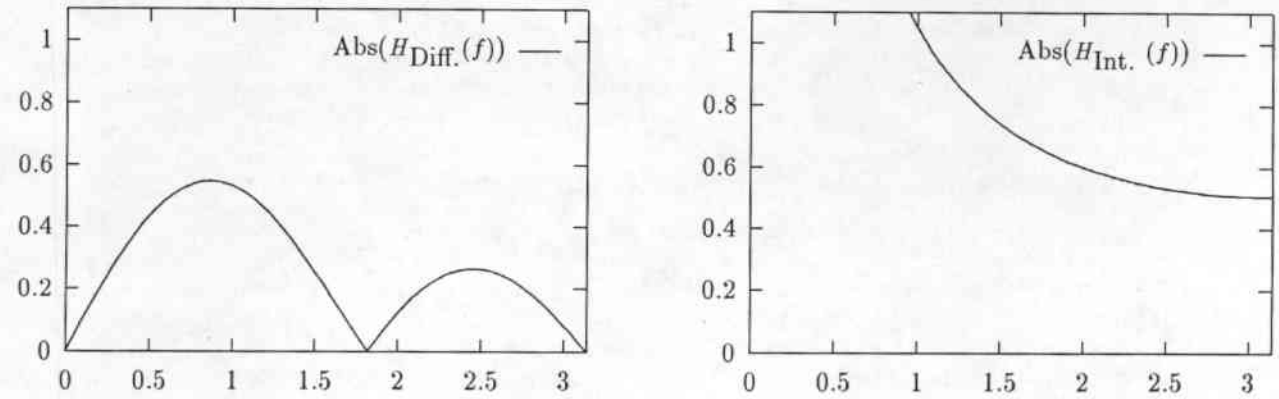


Abbildung 9: Verstärkung bei Differentiation und Integration

Die Kombination beider ergibt eine Bandpaß-Filterung 4. Ordnung des Auditiven Spektrums mit der Transferfunktion $H_4(z)$:

$$H_4(z) = H_{\text{Diff.}}(z)\, H_{\text{Int.}}(z) = 0.1\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{16}$$

Die Eigenschaften dieses Filters lassen sich dem folgenden Beispiel entnehmen:

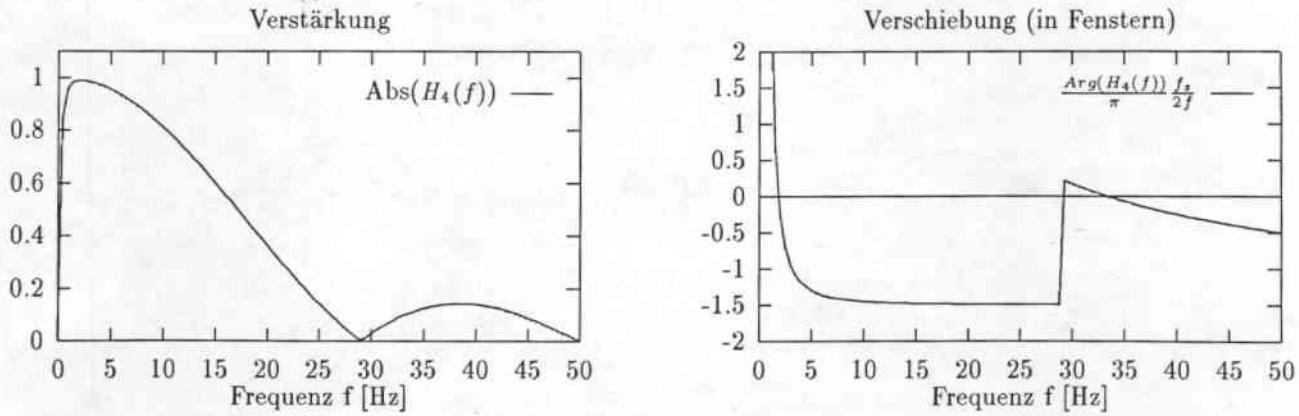


Abbildung 10: Verhalten des RASTA-Bandpaßfilterfilters

Bei der Vorverabeitung im Institut liegen $10ms$ zwischen zwei berechneten Analysefenstern, somit werden 100 Auditive Spektren pro Sekunde ermittelt. Die Bandpaß-Filterung 4. Ordnung hat dann folgende Eigenschaften:

Verstärkung. Der Filter hat scharfe Nullstellen bei 0 Hz und bei 29.1 Hz sowie eine Nullstelle bei 50 Hz. Durch die sehr scharfe Nullstelle bei 0 Hz werden die konstanten und nur sehr langsam variierenden Anteile unterdrückt. Die weniger scharfe Nullstelle bei 50 Hz soll bei der numerischen Fouriertransformation entstehenden Störungen hoher Frequenzen ausblenden.

Der Zweck der scharfen Nullstelle bei 29.1 Hz wird in [3] nicht erläutert. Möglicherweise resultiert sie aus der effizienten Implementierung eines Filters mit den gewünschten Eigenschaften bei 0 Hz und 50 Hz. Inwieweit sich die Nullstelle bei 29.1 Hz auf die Erkennungsleistung auswirkt, konnte nicht geklärt werden. Bei der praktischen Erprobung wurde mit unterschiedlichen Filtern (siehe nächster Abschnitt) gearbeitet, wobei sich das hier angegebene System als am ehesten geeignet erwies.

Zeitverhalten (Verschiebung und Rauschen). Zur Berechnung des Wertes $\dot{y}_{m,i}$ werden zur Differentiation die Eingabewerte $y_{m,i-j}$ für $j \in \{0, 1, 3, 4\}$ und zur Integration ein alter Ausgabewert $\dot{y}_{m,i-1}$ des Filters benutzt.
Dadurch werden zeitliche Informationen im Auditiven Spektrum abhängig von ihrer Änderungsfrequenz verschoben.[7] In der Abbildung 10 ist diese Verschiebung der Ausgabe gegenüber der Eingabe in Analysefenstern angegeben.[8]
Durch die Rückkopplung besitzt das System Einschwingzeiten, in denen verrauschte Werte ausgegeben werden. Aus diesem Grund wird in [3] angemerkt, daß die Analyse in der Stille vor dem Sprachsignal starten soll, so daß sich das System darauf einschwingen kann.

4.4 Modifikationen von PLP und RASTA-PLP

Für die Bewertung der PLP bzw. RASTA-PLP wurde ein Programm vom ICSI-FTP-Server benutzt, das bereits Modifikationen gegenüber der beschriebenen Berechnung der PLP-Koeffizienten aus [3] enthielt und hier zusätzlich noch um andere Bandpaßfilter erweitert wurde. Diese Modifikationen werden im folgenden beschrieben.

Modifikationen bei der PLP. Im Programm des ICSI-FTP-Servers wird eine modifizierte, weniger breite und damit weniger glättende Kritische-Band-Funktion benutzt, um das Auditive Spektrum zu berechnen:

$$\Psi_P(\Omega) = \begin{cases} 10^{\Omega+0.5} & -1.3 \le \Omega \le -0.5 \\ 1 & -0.5 < \Omega < 0.5 \\ 10^{-2.5(\Omega-0.5)} & 0.5 \le \Omega \le 2.5 \\ 0 & \text{sonst} \end{cases} \qquad (17)$$

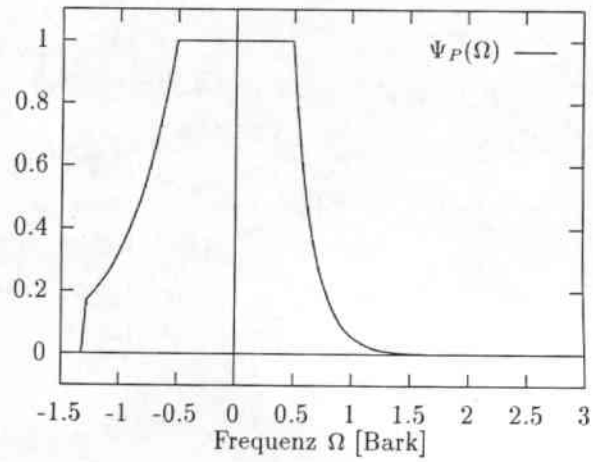


Abbildung 11: Kritische-Band-Funktion im Programm

Außerdem wird in dem Programm eine andere Equal-Loudness-Curve zur Gewichtung des Auditiven Spektrums benutzt:

$$E_P(\omega) = \frac{(\omega^2 + 1.44\ 10^6)\ \omega^4}{(\omega^2 + 1.6\ 10^5)^2\ (\omega^2 + 9.61\ 10^6)}$$

[7] Durch diese Abhängigkeit der Verschiebung von der Änderungsfrequenz ergibt sich das Problem, daß schnelle Änderungen zum Bei bei Plosivlauten einer anderen Verschiebung unterworfen sind als langsamere Änderungen. Durch die Filterung werden neue Phonemüberg (Kontexte) erzeugt (Vokal nach Vokal, Vokal nach Plosivlaut usw.).

[8] In der Spracherkennung werden Systeme anhand von Sprachdaten, in denen Phoneme und ihre Position durch Labels markiert sind, train Um diese Markierung auch bei der RASTA-PLP verwenden zu können, muß die Ausgabe der Koeffizienten in eine Datei entsprechend korri werden. Deshalb kann im Programm zur Implementierung der RASTA-PLP ein Offset angegeben werden, um den die Ausgabe verschoben we soll.

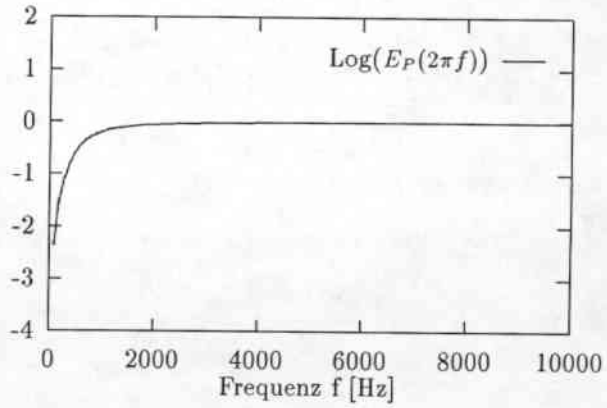


Abbildung 12: Equal-Loudness Kurve im Programm

Modifikationen des Bandpaßfilters. In [3] wird angemerkt, daß neben dem Bandpaß-Filter 4. Ordnung auch andere denkbar seien, die eine scharfe Nullstelle bei 0 Hz aufweisen. Im folgenden sind einige Modifikationen des Filters und die damit bewirkte Änderung der Filtercharakterstik angegeben.

Polstelle. Durch die Angabe der Polstelle bei der Integration, die in [3] bei -0.98 im ICSI-Programm jedoch mit -0.94 angegeben ist, wird die Schärfe der Nullstelle des Bandpaß-Filters bei 0 Hz variiert. Die folgende Abbildung zeigt die Filterfunktion in Abhängigkeit von der Polstelle:

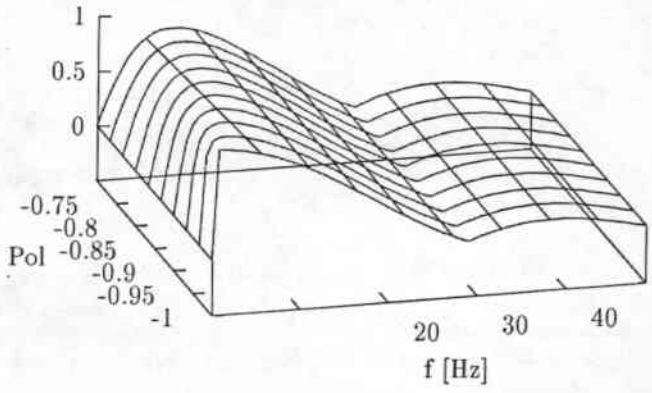


Abbildung 13: Verstärkung des Bandpaßfilters in Abhängigkeit von der Polstelle

Filterordnung. Zur Differentiation können neben dem Filter 4. Ordnung, der in [3] von Hermansky verwendet wird, noch andere differenzierende und integrierende Systeme eingesetzt werden. Der Unterschied zwischen diesen Filtern ist die Regressionsgerade durch die letzten Werte des Auditiven Spektrums. In dem zur Arbeit gehörenden Programm sind Filter der Ordnungen 1 bis 4[9] implementiert.

$$
\begin{aligned}
H_1(z) &= 0.98\ \frac{1-z^{-1}}{1-0.98\ z{-}1} \\
y_1(n) &= 0.98\ (x(n)-x(n-1))\ +\ 0.98\ y(n-1)
\end{aligned}
$$

$$
\begin{aligned}
H_2(z) &= 0.5\ \frac{1-z^{-2}}{1-0.98\ z^{-1}} \\
y_2(n) &= 0.5\ (x(n)-x(n-2)))\ +\ 0.98\ y(n-1)
\end{aligned}
$$

$$
\begin{aligned}
H_3(z) &= 0.142\ \frac{2+z^{-1}-z^{-2}-2\ z^{-3}}{1-0.98\ z^{-1}} \\
y_3(n) &= 0.142\ (2\ x(n)+x(n-1)-x(n-2)-2\ x(n-3))\ +\ 0.98\ y(n-1)
\end{aligned}
\tag{18}
$$

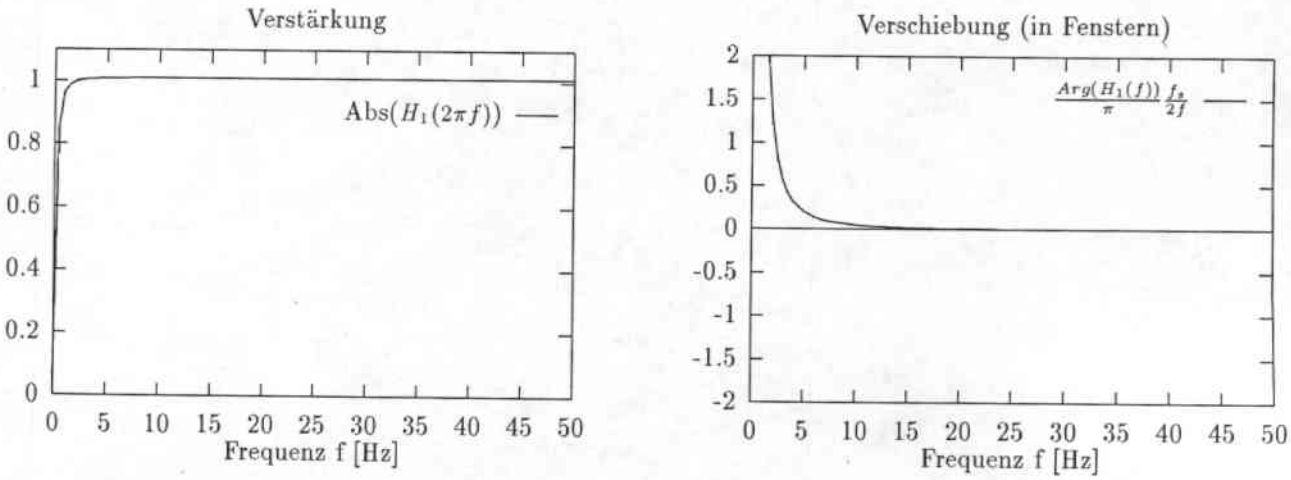


Abbildung 14: Filter 1. Ordnung

Wenn der Filter 1. Ordnung (direkte Differenz der letzten beiden Eingabewerte) mit einem Filterpol von -1 betrieben wird, so entfällt der steile Anstieg der Filterkurve nach 0 Hz. Die Kurve verläuft parallel zur x-Achse. In diesem Fall wird der erste zu filternde Wert einer Folge von allen folgenden abgezogen.
Bezogen auf die Folge der Auditiven Spektren wird der erste ermittelte Vektor des Signals von den anderen substrahiert. Dies Verfahren wird eingesetzt, um das Rauschen in einem Kanal zu eliminieren (blinde Rauschunterdrückung). Dazu sollte im aufgezeichneten Signal vor dem Gesprochenen ein Bereich ohne Sprache enthalten sein.

JRASTA-PLP

Inzwischen wurde eine Erweiterung der RASTA-PLP die sogenannte JRASTA-PLP entwickelt, die zusätzlich auch das additive Rauschen aus dem Kommunikationskanal modelliert. Im Rahmen dieser Arbeit wurde dies noch nicht berücksichtigt.

[9]Der Filter 4. Ordnung (H_4) ist der RASTA-Bandpaßfilter ($H(z)$).

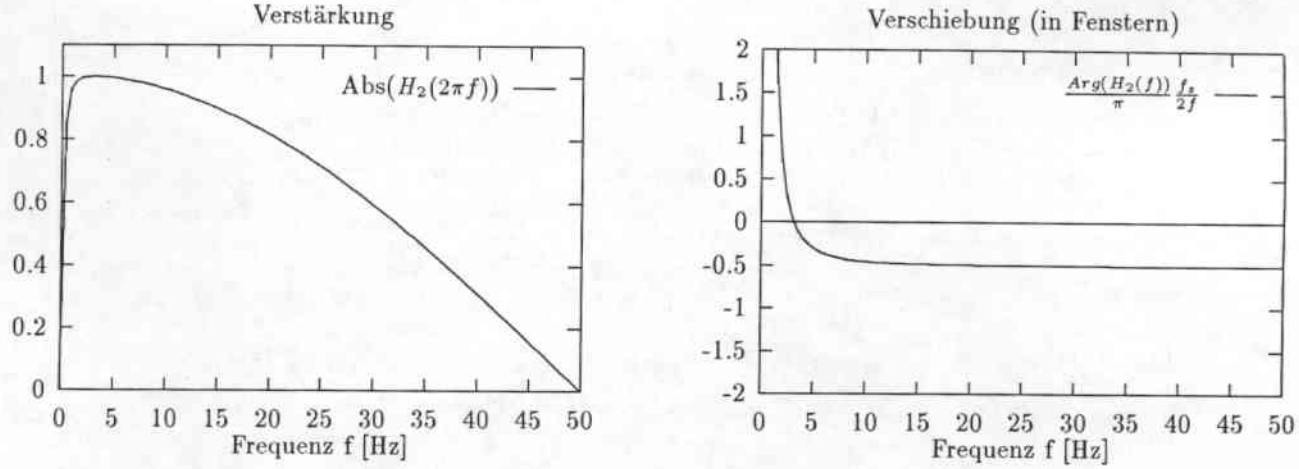


Abbildung 15: Filter 2. Ordnung

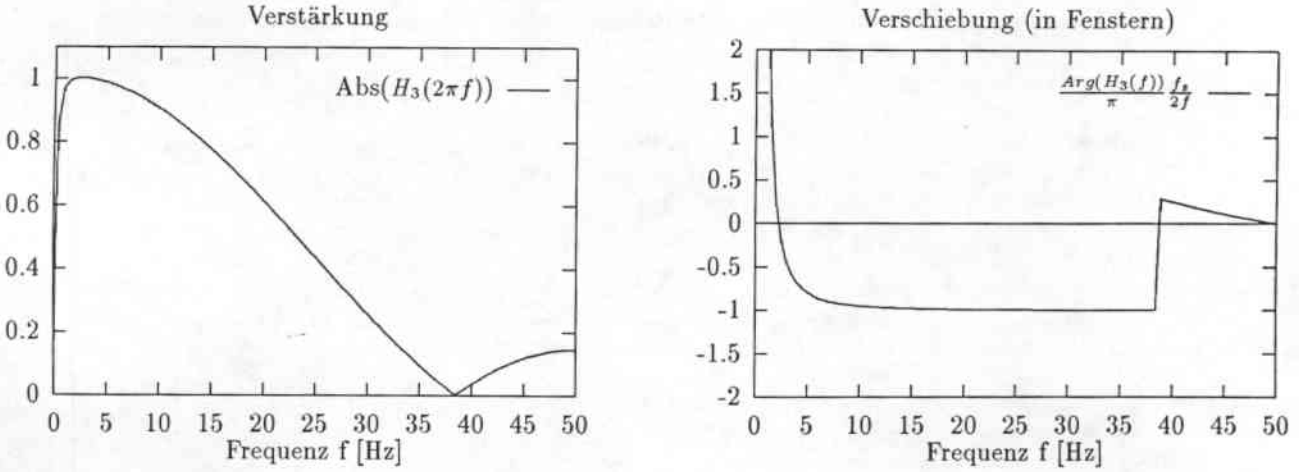


Abbildung 16: Filter 3. Ordnung

Der Unterschied zur RASTA-PLP besteht in der Transformation des auditiven Spektrums vor der Bandpaßfilterung. Je nach Verfahren werden die Koeffizienten

$$\log(\Theta_{m,i}) \qquad \text{RASTA-PLP oder}$$

$$\log(1 + \text{Jah}\ \Theta_{m,i}) \quad \text{JRASTA-PLP}$$

mit dem Bandpaß gefiltert. Die Konstante Jah ist dabei von der Art des Rauschens, das dem Signal im Kommunikationskanal aufaddiert wird, abhängig und muß an die zu erkennenden Sprachdaten angepaßt werden.

5 Praktische Bewertung

5.1 Bewertungskriterien

In dieser Arbeit wurde die praktische Bewertung der vorgestellten Verfahren anhand eines Erkenners auf Basis eines Multi State Time Delayed Neural Networks (MSTDNN siehe [8]) durchgeführt.
Die Datenbasis für die Bewertung besteht aus deutsch buchstabierten Wörtern, die von verschiedenen Sprechern über das normale Telefonnetz (unterschiedliche Telefonverbindungen) aufgenommen und mit der Abtastfrequenz 8kHz digitalisiert wurden. Aus dieser Basis wurden 250 Sätze (buchstabierte Wörter) herangezogen, um den Erkenner zu trainieren. Nach jedem Trainingszyklus wurden die im Mittel auf den Trainingssätzen und weiteren 100 nicht zum Training benutzten Testsätzen erreichten Erkennungsleistungen festgehalten.

5.2 Versuche zur Parametrisierung von PLP und RASTA-PLP

PLP und RASTA-PLP Verfahren sind im Gegensatz zur Melscale-Analyse durch Parameter variierbar. Sie müssen durch Wahl einer geeigneten Parametrisierung auf die jeweilige Erkennungsaufgabe optimiert werden. So verlangen beispielsweise sprecherabhängige und sprecherunabhängige Erkennung unterschiedliche Leistungen der Vorverarbeitung. Im ersten Fall ist es sinnvoll, Merkmale zu extrahieren, die für den Sprecher spezifisch sind. Im zweiten Fall sollten charakteristische Eigenschaften des Sprechers nicht in die extrahierten Merkmale eingehen.
In [2] und [3] werden dazu Parameter vorgeschlagen, jedoch wird auf andere Möglichkeiten hingewiesen. In der Implementierung der PLP und der RASTA-PLP in einem Programm von Chuck Wooters tauchen darüber hinaus weitere Modifikationen auf, deren Einfluß auf die Güte der Verfahren nicht genannt wird. So wurde im Rahmen dieser Arbeit eine Reihe von Versuchen durchgeführt, um für die Gegenüberstellung der Verfahren eine geeignete Parametrisierung zu finden.

Vorrausssetzungen. Um die Anzahl notwendiger Versuche einzuschränken, wurde vorrausgesetzt, daß die Parameter keine großen Wechselwirkungen aufeinander haben, sondern daß jeder Parameter für sich optimiert werden kann.
Die in den Versuchen bestimmten Parameter haben nicht den Anspruch, optimal für die Erkennung von Trainigs- und Testdaten zu sein. Vielmehr sollten sie nahe genug an einer optimalen Lösung liegen, so daß ein qualitativer Vergleich möglich wird.
Für die Bestimmung der Parameter wurde der Erkenner im sogenannten Bootstrapping-Modus betrieben. In dieser Betriebsart benutzt er für alle Trainingssätze (buchstabierte Worte) eine Datei in der die Grenzen der Wörter (Buchstaben) und der darin enthaltenen Phoneme abgelegt sind. Das Bootstrapping dient dazu, das Netz in einem ersten Schritt von Grund auf mit Merkmalsvektoren zu trainieren.
Die riskanteste Vorraussetzung liegt in der jeweiligen Anzahl der Iterationen (Trainingszyklen), die zum Vergleich verschiedener Parameter benutzt wurden. Um bereits nach wenigen Iterationen deutlich schlechtere Parametrisierungen schnell auszusortieren, wurden einige Versuche mit 50 Iterationen durchgeführt. Nach dieser Anzahl sind in der Erkennungsleistung vor allem auf den Testdaten noch große Schwankungen von Iteration zu Iteration zu erkennen. Außerdem kann eine Parametrisierung bei dem verwendeten Erkenner anfangs schlechtere, später aber bessere Leistungen bewirken. Deshalb wurden in Zweifelsfällen bei bestimmten Parameter-Kandidaten weitere 50 oder 150 Iterationen trainiert.

Bewertung der Erkennungsleistung. Die Bewertung stützt sich auf den Durchschnitt der 5 beziehungsweise 30 letzten Iterationen. Bei den 5 letzten Werten erhält man einen Eindruck der zuletzt erreichten Erkennungsleistung, bei den letzten 30 Werten werden langsamere Schwankungen in der Erkennungsleistung abgedämpft. Der Eindruck durch die errechneten Zahlen wurde mit dem Verlauf der Erkennungsleistungen über die Iterationen graphisch überprüft.

5.2.1 Modellordnung p des linearen Prädiktors bei der PLP

In den ersten 50 Iterationen erreichte der Erkenner unter Verwendung von PLP-Koeffizienten der verschiedenen Ordnungen p folgende Erkennungsleistungen:

Ordnung p	Training		Test	
	Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
7	64.54	66.71	61.52	62.93
8	64.92	67.24	62.38	65.03
9	65.18	67.41	62.47	64.57
10	65.42	68.22	62.40	63.30
11	65.91	68.36	62.16	63.37
12	65.33	67.88	62.28	64.30
13	66.00	69.10	62.46	62.70
14	65.75	68.19	62.48	64.73
15	65.58	68.36	61.74	63.36

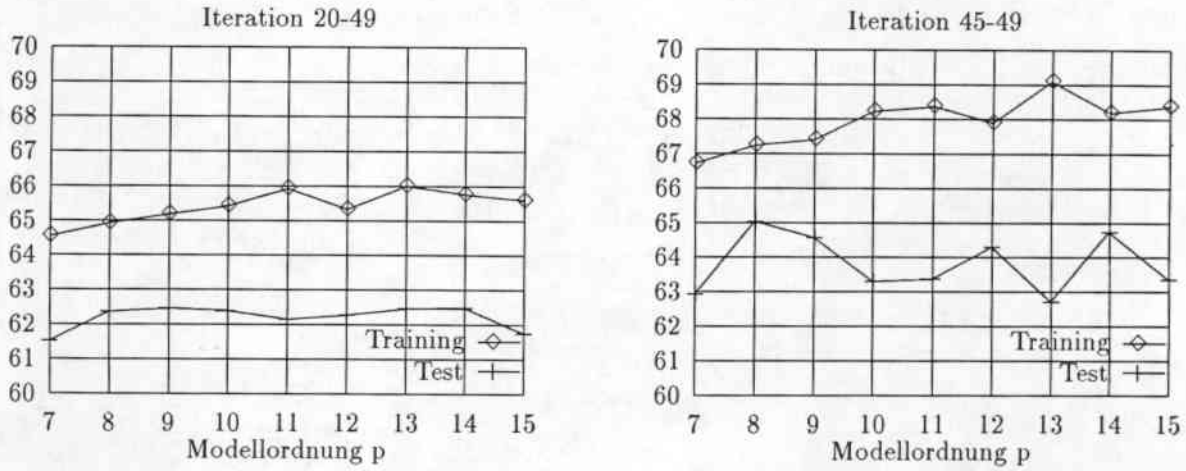


Abbildung 17: Vergleich der Modellordnungen nach 50 Iterationen

Der beste Kandidat für das Training ($p = 13$) und der beste Kandidat für die Testdaten ($p = 8$), der auch in [2] empfohlen wurde, wurden über insgesamt 200 Iterationen trainiert:

Ordnung p	Training		Test	
	Iteration 150-199	Iteration 195-49	Iteration 150-199	Iteration 195-199
8	76.58	77.10	67.14	66.77
13	78.72	79.15	68.92	68.67

Nach einer längeren Trainigszeit erreicht der Erkenner bei Verwendung der Modellordnung $p = 13$ auf Test- und auf Trainingsdaten deutlich bessere Erkennungsleistungen als bei Verwendung von $p = 8$.

5.2.2 Equal-Loudness-Curve

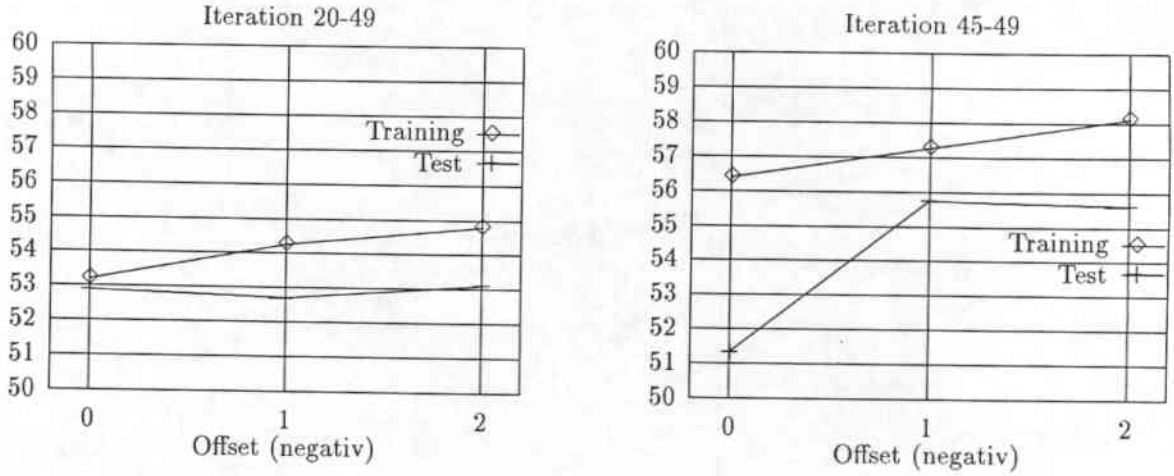
In [2] wird eine andere Equal-Loudness-Curve angegeben, als im Programm vom ICSI-FTP-Server zur Berechnung der PLP- und RASTA-PLP-Koeffizienten benutzt wird.
Bei Versuchen mit RASTA-PLP-Koeffizienten ($p = 15$, Filterpol -0.94 und Offset 0 (siehe 5.2.3)) wurden die Erkennungsleistungen bei Verwendung der Kurven aus [2] und aus dem Programm ermittelt.

ELC aus	Training		Test	
	Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
Programm	54.24	57.25	52.69	55.73
[2]	50.71	53.65	50.74	53.33

5.2.3 Verschiebung der RASTA-Ausgabe.

Durch das Zeitverhalten des Bandpaßfilters tauchen Informationen über gesprochene Phoneme erst verzögert in der Ausgabe auf. Bei Erkennern, die auf markierten Sprachdaten trainiert werden, stimmen dann Markierung und Informationen in der Sprachdarstellung nicht mehr überein. Durch Angabe eines negativen Offsets, um den die Ausgabe verschoben werden soll, läßt sich diese Verschiebung teilweise kompensieren. So ist der Abbildung 10 auf Seite 15 zum Zeitverhalten des Filters 4. Ordnung eine mittlere Verzögerung von ca. 1.5 Fenstern zu entnehmen.
Bei der Modellordnung $p = 15$ und einem Bandpaßfilterpol von -0.98 ergaben die verschiedenen Offsets folgende Erkennungsleistungen:

Offset	Training		Test	
	Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
0	53.20	56.39	52.87	51.32
-1	54.24	57.25	52.69	55.73
-2	54.79	58.11	53.09	55.60



Mit einem Offset von -1 beziehungsweise -2 wurden deutlich bessere Ergebnisse erzielt als ohne Offset. Aus der Betrachtung des Filter-Zeitverhaltens konnte die bessere Erkennungsleistung bei Offset -2 gegenüber Offset -1 nicht vorhergesagt werden.

5.2.4 Bandpaßfilterung vor und nach der Faltung mit der kritischen Band-Funktion

Bei der RASTA-PLP wird das Auditive Spektrum durch einen Bandpaß gefiltert, um Eigenschaften des Kommunikationskanals herauszufiltern. In einem Erkennungsversuch mit Ordnung $p = 15$, Filterpol -0.94 und Offset -1 wurde anstelle des Auditiven Spektrums das Leistungsspektrum gefiltert und erst anschließend mit der Kritischen Bandfunktion gefaltet:

Gefiltertes Spektrum	Training		Test	
	Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
Auditives Spektrum	54.24	57.25	52.69	55,73
Leistungs-Spektrum	54.79	57.68	52.72	55.89

Der nur geringe Unterschied zwischen den beiden Versuchen weist darauf hin, daß nur ein sehr kleiner Anteil an Informationen durch die Filterung des Auditiven Spektrums verloren geht. Die geringfügige Verbesserung wird durch die Filterung von wesentlich mehr Koeffizienten erkauft.

5.2.5 Bandpaß-Filterpol

In [3] wird der Filterpol -0.98 genannt, während im Programm zur Berechnung der RASTA-PLP -0.94 als Vorgabewert für die Berechnung benutzt wird. Mit der Modellordnung $p = 15$ und Offset -1 ergeben sich folgende Erkennungsleistungen:

Filterpol	Training		Test	
	Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
-0.94 (Programm)	54.24	57.25	52.69	55,73
-0.98 ([2])	59.39	62.58	58.10	61.40

5.2.6 Bandpaß-Filterordnung

Es wurden RASTA-Bandpaßfilterordnungen von 1 bis 4 bei dem Filterpol -0.98 und der Modellordnung $p = 14$ für den linearen Prädiktor untersucht. Dabei wurde jeweils ein Offset benutzt, der aus dem zeitlichen Verhalten (siehe Abbildungen 14, 15 und 16 ab Seite 18) ermittelt wurde.

Filterordnung	Offset	Training		Test	
		Iteration 20-49	Iteration 45-49	Iteration 20-49	Iteration 45-49
1	0	56.93	59.98	56.60	58.47
2	-1	58.24	61.46	58.39	60.43
3	-1	58.99	62.48	58.48	60.93
4	-2	59.96	63.14	58.82	60.73

5.3 Versuche zum Vergleich der Verfahren

Anhand der im vorigen Abschnitt angegebenen Versuche wurden folgende Parameter für das PLP beziehungsweise RASTA-PLP-Verfahren zu den abschließenden Vergleichen bestimmt:

- Modellordnung für den linearen Prädiktor $p = 13$ (PLP und RASTA-PLP)
- Equal-Loudness-Curve aus dem Programm von Chuck Wooters (PLP und RASTA-PLP)
- Bandpaß-Filterpol -0.98 (RASTA-PLP)
- Bandpaß-Filterordnung 4 (RASTA-PLP)
- Verschiebung (Offset) -2 (RASTA-PLP)
- Filtern des auditiven Spektrums (RASTA-PLP)

Training mit festen Wort und Phonemgrenzen (Bootstrapping). Der Erkenner wurde in jeweils 200 Iterationen auf Melscale-, PLP- und RASTA-PLP-Koeffizienten trainiert und erreichte damit folgende Erkennungsleistungen:

Verfahren	Training		Test	
	Iteration 170-199	Iteration 195-199	Iteration 170-199	Iteration 195-199
Melscale	81.71	81.94	72.63	73.21
PLP	78.72	79.15	68.92	68.67
RASTA-PLP	74.52	74.83	67.25	67.96

Mit Melscale-Koeffizienten werden die besten Erkennungsleistungen auf den Trainungs- und den Testdaten der gegebenen Datenbasis erreicht, dann folgen PLP- und erst an letzter Stelle RASTA-PLP-Koeffizienten. Die Unterschiede zwischen den Verfahren lassen sich den Abbildungen 18 und 19 [10] entnehmen.

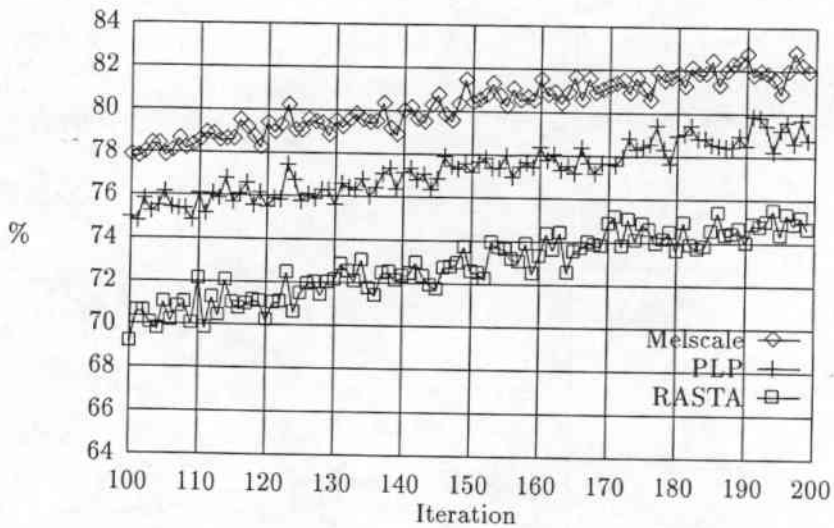


Abbildung 18: Erkennungsleistung auf Trainingsdaten (Bootstrapping)

[10] Zur besseren graphischen Darstellung der Erkennungsleistung auf den Testdaten (Abbildung 19) über die Iterationen wurden die Werte geglättet durch Low-Pass-Filterung mit

$$y(n) = \frac{x(n) + x(n-1)}{4} + \frac{y(n-1)}{2} \quad \text{bzw.} \quad H(z) = \frac{1 + z^{-1}}{4\,(1 - \frac{1}{2}z^{-1})}$$

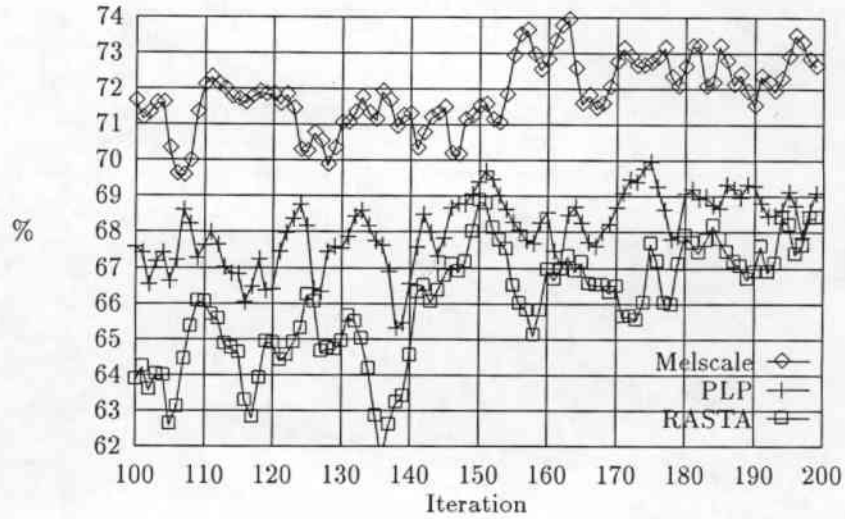


Abbildung 19: Erkennungsleistung (geglättet) auf Testdaten (Bootstrapping)

Training mit variablen Phonemgrenzen innerhalb von Worten. Mit den im Bootstrapping ermittelten Gewichten wurden die Erkenner weitere 200 Iterationen mit variablen Phonemgrenzen trainiert. In dieser Betriebsart benutzt der Erkenner nur die vorgegebenen Wortgrenzen. Die Phonemgrenzen innerhalb eines Wortes kann er variabel bestimmen. Gerade bei Vorverarbeitungsverfahren wie der RASTA-PLP, bei der zeitliche Informationen über gesprochene Phoneme je nach Kontext versetzt auftreten können, ist durch das Training mit variablen Phonemgrenzen eine Steigerung der Erkennungsleistung zu erwarten.

Verfahren	Training		Test	
	Iteration 370-199	Iteration 395-199	Iteration 370-399	Iteration 395-399
Melscale	97.74	97.98	79.19	78.61
PLP	97.68	97.90	74.62	74.36
RASTA-PLP	97.49	97.76	73.86	73.53

Es ist eine deutliche Verbesserung vor allem bei RASTA-PLP und PLP zu erkennen. Bei den Trainingsdatensätzen ergeben sich kaum noch Unterschiede in der Erkennungsleistung bei Verwendung von Melscale-, PLP- und RASTA-PLP-Koeffizienten. Die beiden Abbildungen 20 und 21 zeigen die Verbesserungen, die dabei auftraten.
Auffällig ist, daß der Erkenner bei PLP und RASTA-PLP Koeffizienten schon sehr schnell übertrainiert wird. Bei den Testdaten erreicht das Netz schon nach wenigen (zusätzlichen) Iterationen ein Maximum in der Erkennungsleistung. Nach weiteren Iterationen sinkt diese Leistung herab. Das Netz stellt sich zu sehr auf die Daten für das Training ein.

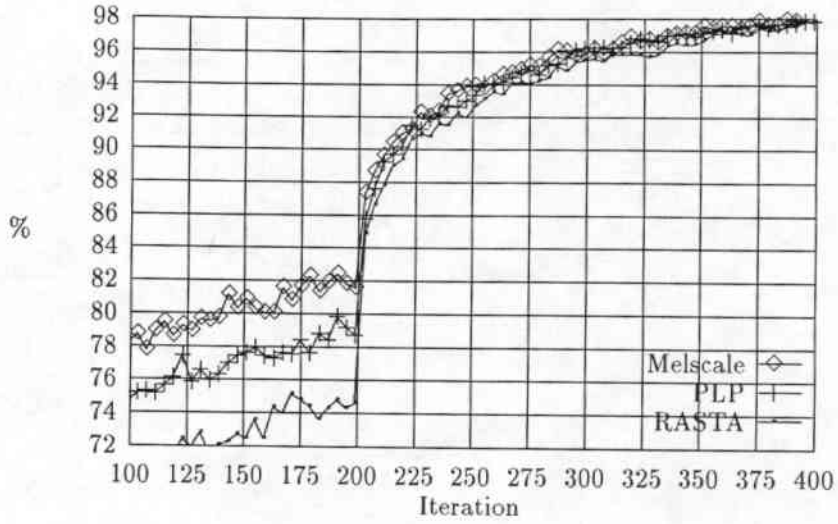


Abbildung 20: Erkennungsleistung auf Trainingsdaten (variable Phonemgrenzen)

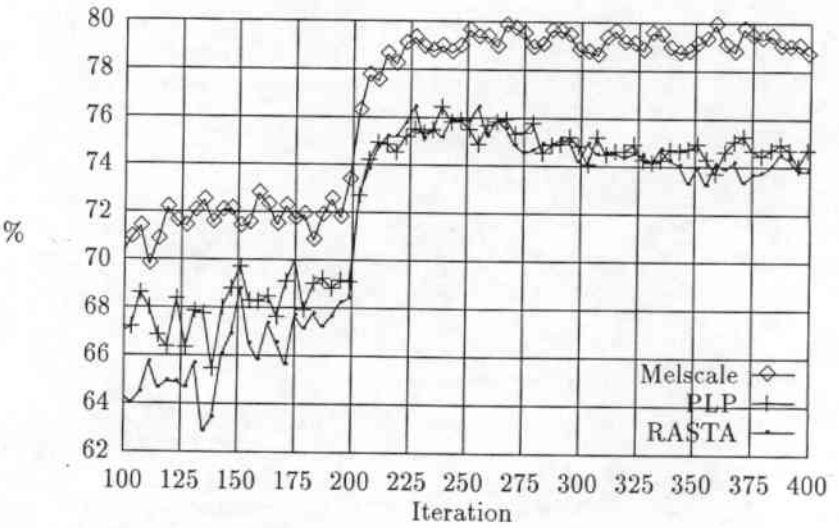


Abbildung 21: Erkennungsleistung (geglättet) auf Testdaten (variable Phonemgrenzen)

5.4 Bewertung

In allen Versuchen ergab sich die gleiche Reihenfolge bei den erreichten Erkennungsleistungen. Bei Verwendung von Melscale Koeffizienten wurden die besten Ergebnisse erreicht, dann folgten PLP und danach RASTA-PLP Koeffizienten.

Für den vorliegenden Buchstabenerkenner auf Basis eines MSTDNNs und die Datenbasis eignen sich PLP- und RASTA-PLP- offensichtlich nicht so gut wie Melscale-Koeffizienten. Die Gründe dafür liegen in folgenden Eigenschaften von PLP und RASTA-PLP:

- Im Gegensatz zur Melscale-Analyse werden bei der Berechnung des Auditiven Spektrums für PLP und RASTA-PLP Frequenzbereiche zum Teil stark überlappend mit der Kritischen-Bandfunktion gefaltet. Die Kritische-Bandfunktion ist breiter als das bei der Melscale-Analyse verwendete Trapez. Dadurch wird das Leistungsspektrum bei PLP und RASTA-PLP stärker geglättet und die Frequenzauflösung geringer als bei der Berechnung von Melscale Koeffizienten.
- Prädiktionsverfahren modellieren die Spracherzeugung beim Menschen. Die zusätzliche Filterung des Signals im Telefonnetz und additives Rauschen beeinträchtigen diese Modellierung empfindlich. In einer bei dieser Arbeit nicht berücksichtigten früheren Versuchsreihe wurde auf Sprachdaten trainiert, die mit einem guten Mikrofon aufgenommen und direkt mit einer Abtastfrequenz von 16 kHz digitalisiert worden waren. Bereits im Bootstrapping erreichte der Erkenner bei Verwendung von PLP-Koeffizienten fast die gleiche Erkennungsleistung wie bei der Verwendung von Melscale Koeffizienten.
- In [5] wird ein Problem bei der Benutzung des RASTA-PLP-Verfahrens genannt: Durch die frequenzabhängige zeitliche Verschleifung von Informationen über gesprochene Phoneme wird ein Kontext erzeugt, der im Sprachsignal nicht vorhanden ist. Zu demselben gesprochenen Laut können je nach Vorgängerlaut verschiedene Merkmalvektoren ermittelt werden. In [5] werden Erkenner mit Ganzwortmodellen, phonembasierte Erkenner mit Triphonemen oder mit einem breiten Eingabekontext empfohlen.
  Der MSTDNN-Erkenner wurde für die Erkennung mit Melscale-Koeffizienten konzipiert. so daß der berücksichtigte Kontext eventuell nicht breit genug für die Verwendung von RASTA-PLP-Koeffizienten ist.
- Wenn vor dem buchstabierten Wort Störungen im Silence vorhanden sind, so wirken sich diese wieder auf die ersten extrahierten Merkmale aus.
- Der MSTDNN-Erkenner wurde auf markierten Phonemgrenzen, wie sie im Sprachsignal vorkommen, trainiert. Da die zeitliche Verzögerung, mit der linguistische Merkmale in die RASTA-PLP-Koeffizienten eingehen, abhängig ist von der Modulationsfrequenz, bewirkt ein fester Offset (hier -2), daß nur ein Teil der Phonem-Merkmale zeitgleich mit der Markierung der Phoneme ist. Im Training führt das in ungünstigen Fällen zu Inkonsistenzen. Die Verbesserung der Erkennungsleistung im Training mit variablen Phonemgrenzen in Vergleich zu fixen Phonemgrenzen kann damit erklärt werden.

A Programmbeschreibung

Programm zur Berechnung von PLP und RASTA-PLP

Das Programm `rastaplp` basiert auf der Implementierung der RASTA-PLP von Chuck Wooters. Neuere Versionen des Programmes befinden sich zur Zeit auf dem FTP-Server icsi@berkeley.edu. Geändert wurden Ein- und Ausgabeformate und die Modulstruktur. Darüberhinaus wurden unterschiedliche Filter und Equal-Loudness-Kurven realisiert.

Aufruf

```
rastaplp -i Eingabe -o Ausgabe [OPTIONEN]
```

Programmablauf

Das Programm berechnet aus den digitalisierten Sprachdaten in der ADC-Datei **Eingabe** eine Folge von Vektoren mit RASTA-PLP- beziehungsweise PLP-Koeffizienten. Die Vektoren werden in der Datei **Ausgabe** in einem am Institut verwendeten maschinenunabhängigen Format abgespeichert.

Optionen

Optionen ohne Argumente ändern den Berechnungs- oder Ausgabemodus.

-P
: Diese Option ist zu setzen, wenn PLP- anstelle von RASTA-PLP-Koeffizienten berechnet werden sollen.

-g
: Bei Angabe dieser Option, wird der Faktor Gain nicht im Ausgabevektor abgespeichert. Der erste Koeffizient in jedem Ausgabevektor ist dann 0.0.

-B
: Bei Angabe dieser Option wird die Ausgabedatei in 1 Byte-Gleitkommazahlen ausgegeben.

Optionen mit Argumenten ändern die in eckigen Klammern angegebenen Vorgabewerte.

-f Abtastfrequenz [16000]
: Angabe der Abtastfrequenz (f_s) in Hertz. Die Frequenz wird zur Berechnung des auditiven Spektrums benötigt.

-w Fensterbreite [16]
: Angabe der Fensterbreite t_f des Analysefensters in Milisekunden. Die voreingestellte Fensterbreite von 20ms ist so gewählt, daß bei einer Abtastfrequenz von 16000 Hz ein Fenster aus 256 Abtastwerten besteht.

-s Schrittweite [10]
: Angabe der Zeit Δt zwischen 2 Analysefenstern in Milisekunden. Bei der vorgegebenen Schrittweite werden 100 Fenster pro Sprechsekunde berechnet.

-L Equal-Loudness-Curve [1]
: Angabe der zu verwendenden Equal-Loudness-Curve E1, E2 oder E3.

-m Linear-Ordnung [14]
: Angabe der Ordnung m des linearen Prädiktors.

-e Spitzenverstärkung [0.6]
: Angabe eines Faktors, der die nichtlineare Verstärkung bei der Berechnung von Cepstral-Koeffizienten aus Prädiktor-Koeffizienten bestimmt.

-r Filterordnung [4]
: Angabe der Filterordnung für den Bandpaßfilter zur Berechnung der RASTA-PLP. Im Programm sind Filter der Ordnungen 1 bis 4 implementiert. (Die charakteristischen Funktionen der Filter sind im Abschnitt 4.3.2 auf Seite 15 angegeben.)

-p Polstelle [-0.98]
: Angabe des Pols für den Bandpaßfilter zur Berechnung der RASTA-PLP. (Die Abhängigkeit der charakteristischen Filterfunktion wird im Abschnitt 4.3.2 auf Seite 15 skizziert.)

-n Vektorbreite [16]
: Angabe der Breite für die Ausgabevektoren.

-d Debugmode [0]
: Auswahl der Informationen, die im Programmablauf über die Standard-Fehlerausgabe ausgegeben werden.

B Literaturverzeichnis

Literatur

[1] R. N. Bracewell. *Schnelle Hartley-Transformation.* Oldenbourg, 1990.

[2] Hynek Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.*, 87(4):pp. 1738–1752, April 1990.

[3] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. RASTA-PLP Speech Analysis Technique. *Proc. ICASSP, IEEE*, September 1992.

[4] Hermann Hild and Alex Waibel. Multi-speaker/speaker-independent architectures for the multi-state time delay neural network. *Proc. ICASSP, IEEE*, April 1993.

[5] Joachim Koehler, Nelson Morgan, H. Guenter Hynek Hermansky Hirsch, and Grace Tong. Integrating RASTA-PLP into Speech Recognition. *Proc. ICASSP, IEEE*, 1994.

[6] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1987.

[7] Bernhard Suhm. Fließbandverarbeitung in der Vorverarbeitung von Sprachsignalen. Studienarbeit, Institut für Logik, Komplexität und Deduktionssysteme, 1992.

[8] Alex Waibel and Kai-Fu Lee, editors. *Readings in Speech Recognition.* Morgan Kaufmann Publishers, 1990.