Schnelle adaptive Sprechernormierung für die Spracherkennung

Studienarbeit

von Kjell Schubert

Institut für Logik, Komplexität und Deduktionssysteme
Prof. Dr. Alex Waibel
Universität Karlsruhe – SS 1998

Zusammenfassung

Verschiedene Arten der Sprechernormierung sind in der automatischen Spracherkennung üblich. Unter ihnen ist die Vokaltraktlängennormierung VTLN dafür zuständig, die schädliche Varianz der Vokaltraktlängen verschiedener Sprecher zu kompensieren. Das bisher an der Universität Karlsruhe eingesetzte Verfahren basiert auf einer Maximum-Likelihood basierten Schätzung der Vokaltraktlänge und ist für den Einsatz in echtzeitfähigen schritthaltenden Erkennern ungeeignet. In der vorliegenden Arbeit wird eine VTLN-Variante vorgestellt, die diesen Nachteil nicht besitzt und bei einer maximalen Verzögerung des Erkenners um 2 Sekunden die Wortfehlerrate um 3.5% reduziert. Diese Reduktion ist nur wenig geringer als diejenige von 5%, die mit der bisher verwendeten Methode erreicht wird.

Inhaltsverzeichnis

1 Einleitung

Eines der aktuellen Probleme beim Entwurf sprecherunabhängiger Spracherkenner ist die Sprecherabhängigkeit des Sprachsignals, welche zu einer schlechteren Erkennungsrate im Vergleich zu sprecherabhängigen Erkennern führt. Die Gründe für diese Sprecherabhängigkeit des Sprachsignals sind vielfältig: Dialekt und Vokaltraktform des Sprechers gehören zu den wichtigsten. Um den Einfluß der Vokaltraktform auf die aus dem Sprachsignal gewonnenen Merkmale zu reduzieren, kann eine sogenannte Vokaltraktlängennormierung (VTLN) auf dem Signal durchgeführt werden. Dazu wird im Spektralbereich eine Transformation der Frequenzachse vorgenommen, um über diesen Weg eine Verringerung der Varianz der Merkmalsvektoren innerhalb jeder Phonemklasse zu erreichen. Die bisher benutzten Verfahren zur Bestimmung der Transformationsparameter waren aber leider zu zeitaufwendig, um in Erkennern eingesetzt zu werden, die in Echtzeit arbeiten sollen. Außerdem mußte ein relativ langer Block aus dem Sprachsignal (bis zu 50sec) benutzt werden, um die Transformationsparameter genau genug schätzen zu können, was den Einsatz der VTLN in Echtzeiterkennern weiter erschwerte, beziehungsweise unmöglich machte. Ziel dieser Studienarbeit war es, Lösungen für diese beiden Probleme zu finden.

2 Einführung in VTLN Verfahren

Einer der wichtigsten Gründe für die Sprecherabhängigkeit des Sprachsignals ist die Vokaltraktlänge des Sprechers. Männer haben eine durchschnittliche Vokaltraktlänge von etwa 18 cm, bei Frauen liegt dieser Wert im Mittel bei 13 cm. Diese Unterschiede in den Vokaltraktlängen der Sprecher führen dazu, daß Formanten, also die Resonanzfrequenzen im Spektrum, bei Frauen durchschnittlich 20% höher liegen, als bei Männern. Diese Verschiebung hat einen negativen Einfluß auf die Erkennungsrate des sprecherunabhängigen Spracherkenners, weil damit die 'Ähnlichkeit' der Spektren und damit der Merkmalsvektoren innerhalb von ein und derselben (Sub-)Phonemklasse verringert wird. Das heißt, die Merkmalsvektoren eines bestimmten Phonems, das von verschiedenen Sprechern stammt, liegen im Merkmalsraum relativ weit auseinander. In Abbildung 1 wird die Formantenverschiebung bei 3 Phonemklassen K_1, K_2 und K_3 dargestellt, wobei auf den Achsen die beiden Formanten F_1 und F_2 aufgetragen sind. Daß durch diese Verschiebung die Klassifizierung erschwert werden kann, zeigen die Phonemklassen K_2 und K_3: wenn nur ein einzelner Sprecher betrachtet wird, überlappen sich die beiden Klassen nicht, und sind somit leicht unterscheidbar. Wenn jedoch mehrere Sprecher betrachtet werden, kommt es zu Überlappungen der beiden Phonemklassen im mit X gekennzeichneten Bereich.

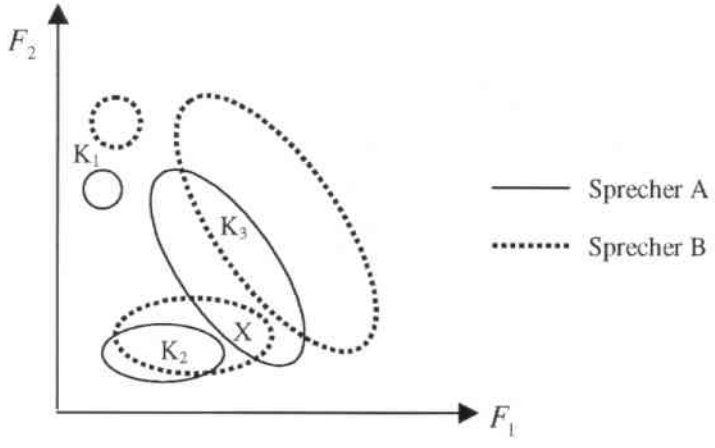

Abbildung 1. Vokaldreiecke zweier Sprecher mit unterschiedlicher Vokaltraktlänge

Um dieses Manko teilweise zu kompensieren, müssen für eine ausreichende Modellierung der Klassenverteilungen im Merkmalsraum im Vergleich zu sprecherabhängigen Erkennern viel mehr Trainingsdaten aufgewendet werden.
Da sprecherunabhängige Erkenner alleine durch Vergrößerung der Trainingsdatenmenge aber noch nicht die Leistung vergleichbarer sprecherabhängiger Erkenner erreichen, ist es sinnvoll die Varianz innerhalb der einzelnen Phonemklassen (bzw. Subphonemklassen) durch eine geeignete Transformation zu reduzieren. Hierzu bietet sich die VTLN an. Diese nutzt die Tatsache aus, daß die

Lage der Resonanzfrequenzen im Spektrum bei ein und demselben Phonem annähernd umgekehrt proportional zur Länge des Vokaltraktes des Sprechers ist. Um mit diesem Wissen die Sprachsignale verschiedener Sprecher zu normieren, wird auf einem Ausschnitt des Sprachsignals ein zur Vokaltraktlänge des Sprechers korrelierter Parameter geschätzt, mit dem das Spektrum gestreckt beziehungsweise gestaucht wird. Dieser Parameter wird folgend in Anlehnung an die englische Literatur Warpingfaktor genannt.
Sowohl für die Schätzung des Warpingfaktors basierend auf dem Sprachsignal, als auch für die Verzerrung des Spektrums existieren mehrere Verfahren. In [4] wurden 2 Klassen von Warpfunktionen getestet: lineare und nichtlineare.

$$f'=k_s f \qquad \text{(linear)}$$
$$f'=k_s^{3f/8000} f \qquad \text{(nichtlinear)}$$

Diese Funktionen wurden in [4] aus einfachen Vokaltraktmodellen abgeleitet: dem 'uniform tube model' und dem Helmholtz-Resonator-Modell. Tests mit einem HMM-Erkenner in diesem Artikel zeigten eine geringfügige Überlegenheit der nichtlinearen Warpfunktion. Mit ihr wurden 57.1% Wortfehlerrate erreicht, gegenüber 57.3% mit linearem Warping. Eine weitere stückweise lineare Warpfunktion wurde in [3] vorgestellt und in [1] mit der nichtlinearen Version verglichen:

$$f'=\begin{cases} \alpha_s^{-1} f & , f<F \\ bf+c & , f\geq F \end{cases}$$

$$f'=\begin{cases} \alpha_s^{-3f/8000} f & f<F \\ bf+c & f\geq F \end{cases}$$

wobei α_s^{-1} der Warpingfaktor für den jeweiligen Sprecher, basierend auf der für ihn geschätzten Vokaltraktlänge, ist. b und c sind Konstanten, die gemäß $\alpha_s F=bF+c$ und $8000b+c=8000$ berechnet werden. F ist die konstante Schranke, an der die 2 linearen Teilstücke der Warpfunktion zusammentreffen, im Test wurde sie auf 5600Hz gesetzt. Im Vergleich mit der nichtlinearen Warpfunktion zeigte die stückweise lineare in [1] bessere Erkennungraten. Aus diesem Grund haben wir im Rahmen dieser Studienarbeit diese Warpfunktion für unsere Experimente gewählt (Abbildung 2).

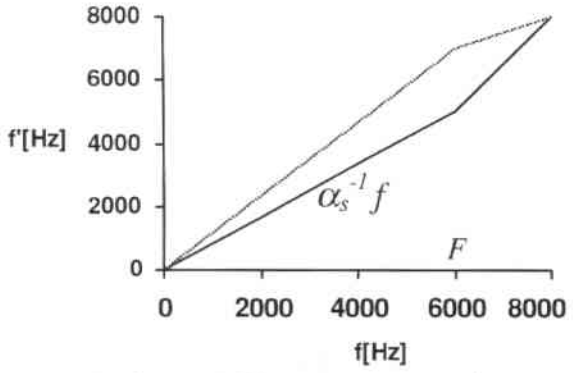

Abbildung 2. Bild der stückweise linearen Warpfunktion für 2 Beispielsprecher

Wie bereits erwähnt, gibt es nicht nur für die Art und Weise der Verzerrung des Spektrums, sondern auch für die Wahl der Warpingparameter verschiedene Methoden. Dabei lassen sich diese Methoden in drei Gruppen einteilen: formantenbasierte, pitchbasierte und ML-basierte (Maximum Likelihood) Verfahren.

Die Methoden aus der erstgenannten Gruppe bestimmen die durchschnittliche Formantenfrequenz f für F1,F2 oder F3 auf einem genügend großen Ausschnitt einer Äußerung (meist mehrere Sätze). Außerdem wird der durchschnittliche Wert f^* für diese Formantenfrequenz über alle Sprecher aus dem Corpus bestimmt. Die Spektren für diese Äußerung, bzw. diesen Sprecher, werden dann genau so verzerrt, daß die betrachtete Formante, die im unverzerrten Spektrum (im Durchschnitt) bei f liegt, im verzerrten Spektrum (im Durchschnitt) mit f^* zusammenfällt. Beispielsweise würde der Warpingfaktor bei Benutzung der linearen Warpingfunktion mit dieser Methode gleich f / f^* gesetzt werden. Genaueres zu dieser Methode wird in [1] beschrieben.

Die pitchbasierten Verfahren machen sich die Korrelation zwischen Vokaltraktlänge und durchschnittlicher Stimmbandfrequenz zunutze: jemand, der eine hohe durchschnittliche Stimmbandfrequenz besitzt, hat im Allgemeinen auch einen kleinen Vokaltrakt. Deshalb bestimmen diese Verfahren, ähnlich wie die formantenbasierten, auf einem genügend großen Ausschnitt einer Äußerung (wieder meist mehrere Sätze) den durchschnittlichen Wert für die Stimmbandfrequenz F0, und ermitteln mit Hilfe einer vorher parametrisierten Abbildung den optimalen Warpingfaktor für diesen F0 Wert. Diese Abbildung wird angelegt, indem auf der gesamten Trainingsmenge für jeden Sprecher der durchschnittliche F0 Wert bestimmt wird, und dazu der Warpingfaktor, der für diesen Sprecher die beste Erkennungsrate gebracht hat. Aus dieser Information lassen sich die Parameter der Abbildung durch ein geeignetes Verfahren (z.B. Regression) bestimmen. Eine detailliertere Beschreibung dieser Methode kann [5] entnommen werden.

Die dritte und letzte Gruppe bilden die ML-basierten Verfahren. Diese Verfahren können im Gegensatz zu den beiden erstgenannten nur mit Hilfe des akusti-

schen Modells Λ des Erkenners benutzt werden. Es wird derjenige Faktor α_s zum Strecken des Spektrums benutzt, der die Wahrscheinlichkeit für die Beobachtung dieser Äußerung maximiert:

$$\alpha_s^* = argmax_\alpha P(X(\alpha)/\Lambda, W)$$

wobei $X(\alpha)$ die Folge der Merkmalsvektoren der Äußerung ist, deren Spektren mit Faktor α verzerrt wurden. W ist die zugehörige Transkription.

Vergleiche von Formanten und ML-Verfahren finden sich in [1], wo alternativ die Formanten F1, F2 und F3 zur Bestimmung des Warpparameters benutzt wurden (Tabelle 1). In [5] wurden pitchbasierte und ML-Verfahren verglichen (Tabelle 2). Dabei hat sich eine Überlegenheit der ML-Verfahren gegenüber pitch- und formantenbasierten Verfahren gezeigt. Der Einsatz von ML-Verfahren in Echtzeiterkennern scheiterte jedoch bis jetzt an ihrem hohen Zeitaufwand.

Modes	Basissystem	F1	F2	F3	ML
Linear	21.8%	20.5%	21.9%	21.6%	19.8%
Nonlinear	21.8%	21.5%	22.7%	21.6%	21.0%

Tabelle 1.Wortfehlerraten von formanten- und ML basierten VTLN [1]

Basissystem	F0	ML
26.1%	24.4%	24.0%

Tabelle 2. Wortfehlerraten von pitch- und ML basierten Verfahren [5]

3 Integration der VTLN in die Vorverarbeitung

Ziel der Vorverarbeitung ist es, das analoge Sprachsignal in eine Folge von Merkmalsvektoren zu transformieren. Diese Merkmalsvektoren sollen möglichst alle für die Spracherkennung relevanten Informationen enthalten, und dennoch niedrigdimensional genug sein, um die Erkennung in vertretbarer Zeit durchführen zu können.
Die Vorverarbeitungsstufe erhält als Eingabe ein analoges Sprachsignal. Dieses Signal muß zunächst zu äquidistanten Zeitpunkten abgetastet werden ('sampling'), was eine Folge von Amplitudenwerten liefert. Nach dem Nyquist-Theorem kann aus dieser Folge das analoge Signal reproduziert werden, wenn die Abtastrate mindestens doppelt so groß ist, wie die größte im analogen Signal vorkommende Frequenz. Da Frequenzanteile, die über der doppelten Abtastfrequenz liegen, im Signal für Verzerrungen sorgen ('aliasing'), muß vor der Abtastung noch eine Tiefpaßfilterung mit dieser Grenzfrequenz durchgeführt werden.
Aus der so erhaltenen Amplitudenwertfolge werden durch Multiplikation mit einer geeigneten Fensterfunktion alle 10 ms Blöcke von etwa 15 bis 20 ms Länge ausgeschnitten. Bei den gebräuchlichen Abtastraten von 8 bis 20 kHz entspricht das meistens einer Blockgröße von 256 Amplitudenwerten. Für die Weiterverarbeitung dieser Blöcke ('frames') existieren viele verschiedene Methoden, wobei Fouriertransformation und Lineare Prediktion zu den meistbenutzten zählen. Im Rahmen dieser Studienarbeit haben wir eine Transformation gewählt, die uns 13 sogenannte Melscaled Frequency Cepstral Coefficients (MFCC) liefert. Diese erhält man, indem man das durch die Fouriertransformation erhaltene Betragsspektrum gemäß einer Melscale nichtlinear verzerrt, komprimiert, logarithmiert, und anschließend einer Cosinustransformation unterzieht. Anschließend wurden diese MFCC noch durch ihre Deltakoeffizienten, Deltadeltakoeffizienten und Energiemessung ergänzt und ergaben einen 40-elementigen Vektor. In einem letzten Schritt haben wir diesen Vektor mittels einer LDA-Transformation auf einen 28-elementigen Merkmalsvektor reduziert, der schließlich das Ergebnis der Vorverarbeitungsstufe darstellt.

Abbildung 3. Vorverarbeitung

Die VTLN wird, wie in der Einleitung bereits erwähnt wurde, durch eine Transformation der Frequenzachse vorgenommen. Dabei gibt es mehrere Möglichkeiten diesen Schritt in die Vorverarbeitung zu integrieren:

1. Transformation des Spektrums direkt nach der Berechnung der Fouriertransformation. Diese Vorgehensweise ist in JANUS[1] implementiert, und wird auch in [3] vorgeschlagen.
2. Modifikation der Abstände und der Breite der MEL-Filterbänke wird in [2] vorgeschlagen. Vorteil gegenüber der erstgenannten Methode ist, daß der Rechenaufwand geringfügig verringert wird, wobei aber der Rechenaufwand für die Transformation schon in der ersten Methode im allgemeinen vernachlässigbar klein ist.
3. Abtastratenwandlung des Signals im Zeitbereich. Nachteile dieser Methode sind, daß damit nur lineare Transformationen der Frequenzachse durchgeführt werden können, sowie der erhöhte Rechenaufwand. Deshalb ist diese Methode in der Praxis nicht zu empfehlen.

In der unteren Abbildung wird die im JANUS verwendete Methode der Integration der VTLN in die Vorverarbeitung dargestellt.

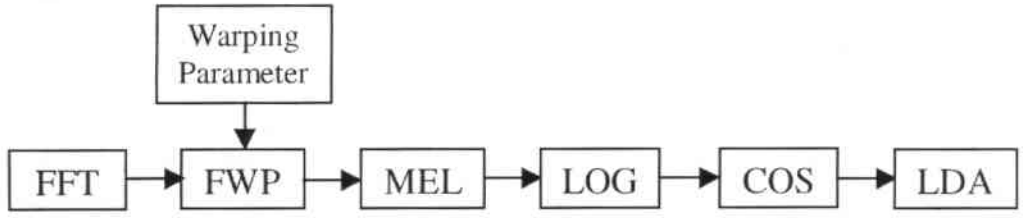

Abbildung 4. Vorverarbeitung mit VTLN (FrequencyWarPing) nach Methode 1

[1] Siehe Anhang A.1

4 HMM Training mit ML-VTLN

Ziel der Trainingsprozedur mit VTLN ist es, die HMM Parameter über dem sprechernormierten Merkmalsraum zu bestimmen. Während des Trainings müssen jetzt nicht nur wie bei der Standardtrainingsprozedur ohne VTLN die Emissionswahrscheinlichkeiten der HMM-Zustände bestimmt werden, sondern zusätzlich noch der (im ML-Sinne) optimale Warpingfaktor für jeden Sprecher. Die so erweiterte Trainingsprozedur läuft folgendermaßen ab:

1. Initialisiere den Warpingfaktor für jeden Sprecher mit 1.0
2. Führe die Standard HMM-Trainingsprozedur (EM-Algorithmus) basierend auf den aktuellen Warpingfaktoren durch (z.B. mit Viterbi)
3. Ersetze den Warpingfaktor für jeden Sprecher jetzt durch den Faktor, der Gleichung
   (4.1) $\alpha_s^* = argmax_\alpha P(X(\alpha)/\Lambda,W)$
   erfüllt. W ist die Transkription des Sprachsignals X, und $X(\alpha)$ ist das mit α gewarpte Signal.
4. Solange es signifikante Änderungen in den Warpingfaktoren gibt, gehe zu Schritt 2

Hauptproblem in dieser Prozedur ist die Lösung der Gleichung (4.1) im dritten Schritt. Da für die Lösung dieser Gleichung im allgemeinen keine geschlossene Form angegeben werden kann, wird α_s^* durch Suche auf einem Raster bestimmt. In [2] wurde ein Raster vorgeschlagen, das 13 Werte im Bereich von 0.88 bis 1.12 im Abstand von 0.02 enthält. Dieser Bereich spiegelt die erwarte 25% Schwankung in der Vokaltraktlänge von Männern und Frauen wider. Die Rastersuche in Schritt 3 sieht dann in der Praxis so aus, daß die Merkmalsvektoren für jeden der 13 Warpingfaktoren im Raster bestimmt werden, und der ML-Score[1] entlang der schon für das HMM-Training berechneten Pfade berechnet wird.
Zu beachten ist noch, daß der Score nur auf den stimmhaften Phonemen bestimmt wird, weil dadurch eine bessere Erkennungsrate erreicht wird, also $\alpha_s^* = argmax_\alpha P(X(\alpha)/\Lambda,W,\sigma)$, wobei σ der Pfad ist. Eine Erklärung für diese Verbesserung ist, daß die Vokaltraktlänge sich hauptsächlich auf stimmhafte Laute, insbesondere Vokale, auswirkt und weniger auf stimmlose Phoneme wie Zischlaute. Die Unterscheidung zwischen stimmhaft und stimmlos ist sicher nicht optimal, wenn man bedenkt, daß auch Zischlaute stimmhaft ausgesprochen werden können und zum Beispiel das Phonem 'h' vor Vokalen zwar stimmlos ausgesprochen wird, aber trotzdem die Formantenstruktur des Folgevokals zeigt.

[1] $-\log P(X/\Lambda)$

Trotzdem ist dieses Unterscheidungskriterium gut genug, um eine Verbesserung der Erkennungsleistung zu bringen.

5 Erkennung mit ML-VTLN

Die VTLN-Erkennungsprozedur soll wie auch die ML-Erkennungsprozedur ohne VTLN die wahrscheinlichste Wortfolge für das gegebene Sprachsignal bestimmen. Bei Einsatz von VTLN muß zuvor allerdings der Warpingparameter des Sprechers geschätzt werden. Wie bereits im vorigen Kapitel erwähnt, wird derjenige Faktor α_s^* zum Strecken des Spektrums benutzt, der die Wahrscheinlichkeit für die Beobachtung dieser Äußerung maximiert:

$$(5.1)\ \alpha_s^* = argmax_\alpha P(X(\alpha)|\Lambda, H(\alpha))$$

Die Lösung dieser Gleichung muß wie im Training durch eine Rastersuche approximiert werden. Im Gegensatz zum Training haben wir jetzt aber keine Transkription W der zu dekodierenden Äußerung gegeben. Infolgedessen muß während der Rastersuche für jeden Warpingfaktor α die HMM-Erkennungsprozedur auf dem mit diesem Warpingfaktor gestreckten Sprachsignal $X(\alpha)$ durchgeführt werden, um so die vom Warpingfaktor abhängige beste Hypothese $H(\alpha)$ zu finden. Unter den getesteten Warpingfaktoren wird dann der beste gemäß Gleichung (5.1) ausgewählt. Mittels dieses ML-VTLN-Verfahrens konnte die Erkennungsrate des Ausgangserkenners auf dem GSST (siehe Anhang A.1) von 83.9% ohne VTLN auf 85.0% mit VTLN gesteigert werden. Die mit der stückweise linearen Warpingfunktion theoretisch maximal erreichbare Erkennungsrate lag bei 87.7%. Es werden also ca. 30% aller durch diese VTLN-Methode korrigierbaren Fehler korrigiert. Die maximal erreichbare Erkennungsrate bestimmten wir, indem wir für jede Äußerung nicht das α_s^* auswählten, das den besten ML-Score (laut Gleichung 5.1) liefert, sondern dasjenige, mit dem die beste Erkennungsrate erzielt wird.

Basissystem (ohne VTLN)	83.9%
VTLN (13 Hypothesen)	85.0%
Theoretisches Optimum	87.7%

Tabelle 3. Verbesserung der Erkennungsraten mit VTLN

Die Implementierung des ML-VTLN-Verfahrens ist sehr einfach, problematisch ist jedoch der Anstieg des Zeitbedarfs für die Erkennung: wo bei einem vergleichbaren Erkenner ohne VTLN nur eine einzige Hypothese bestimmt werden muß, müssen beim Erkenner mit VTLN und der Benutzung eines Rasters mit üblicherweise 13 Warpingfaktoren auch 13 Hypothesen berechnet werden. Dies ist einer der Gründe dafür, daß dieses Verfahren in zeitkritischen Erkennern noch nicht eingesetzt werden kann.

In [1] konnte das Verfahren jedoch bedeutend beschleunigt werden, wobei die Leistung in bezug auf die Fehlerrate nur unbedeutend verschlechtert wurde. Dabei wird nicht für jeden Warpingfaktor im Raster eine eigene Hypothese $H(\alpha)$ bestimmt, sondern der Erkenner bestimmt eine einzige Hypothese $H(\alpha=1.0)$ ohne Einsatz der VTLN, und bestimmt dann basierend auf dieser Hypothese wie schon im Training den ML-Score für jeden Warpingfaktor. Dadurch spart man sich die Durchführung der zeitaufwendigen Viterbi-Suche für jeden Warpingfaktor, nur die im Verhältnis dazu viel schnellere ML-Scoreberechnung muß noch für jeden einzelnen Faktor durchgeführt werden.

Basissystem (ohne VTLN)	83.9%
Langsame VTLN (13 Hypothesen)	85.0%
Schnellere VTLN (nur noch eine Hypothese)	84.8%

Tabelle 4. Erkennungsraten mit beschleunigter VTLN (mit nur noch einer Hypothese)

5.1 Schnelle Warpingfaktorschätzung mit einfacheren HMMs

Aber auch mit der letztgenannten Abwandlung der Erkennungsprozedur ist der Zeitbedarf des Erkenners mit VTLN noch mehr als zweimal so hoch wie der des gleichen Erkenners ohne VTLN. Deswegen wurde in [2] eine weitere Möglichkeit zur effizienteren ML-Schätzung von Warpingfaktoren vorgestellt: Am Ende der im vorigen Abschnitt angegebenen Trainingsprozedur ist für jeden Sprecher der optimale Warpingfaktor bestimmt worden. Mit Hilfe dieser Information werden für jeden der 13 Faktoren Mixturen aus multivariaten Gaussverteilungen bestimmt, die die Verteilung der 13 Sprecherklassen im (unverzerrten) Merkmalsraum beschreiben. Dazu wird jede Sprecherklasse mit den unverzerrten Trainingsdaten von allen diesem Faktor zugeordneten Sprechern trainiert, so daß damit 13 akustische Modelle $\Lambda'(\alpha)$ erzeugt werden.
Für die Warpingfaktorschätzung laut Gleichung (5.1) wird dann nicht mehr der relativ langsame volle Erkenner eingesetzt, sondern es werden nur noch die im Vergleich dazu viel einfacheren und schnelleren Wahrscheinlichkeitsberechnungen auf den 13 Sprachmodellen $\Lambda'(\alpha)$ durchgeführt. Analog zu Gleichung (5.1) wählt man dann denjenigen Warpingfaktor aus, der $P(X/\Lambda'(\alpha))$ maximiert:

$$(5.2)\quad \alpha_s^* = argmax_\alpha P(X/\Lambda'(\alpha))$$

Im Vergleich zu (5.1) ist auch keine Berechnung einer Hypothese mehr notwendig, da jedes der 13 Sprechermodelle $\Lambda'(\alpha)$ nur noch auf einer einzigen Klasse basiert.

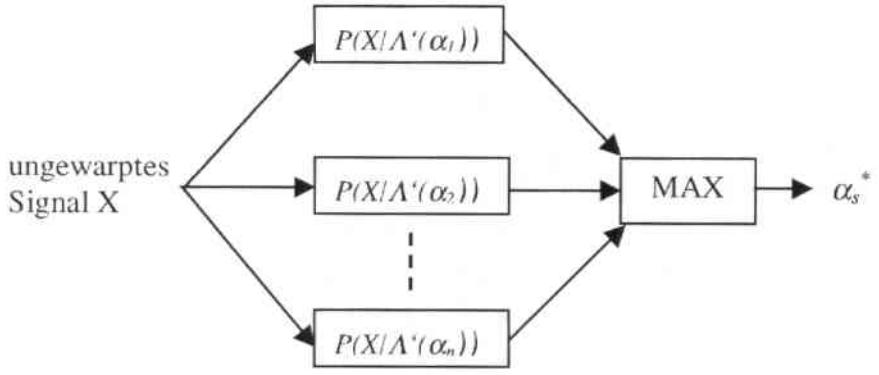

Abbildung 5. Mixturbasierte Warpingfaktorschätzung nach [2]

Analog zum oben dargestellten Warpingfaktorschätzer aus [2] wurden im Rahmen dieser Studienarbeit mehrere alternative Schätzverfahren getestet. Alle diese Schätzer haben den Vorteil, daß sie wie das Verfahren in [2] um ein Vielfaches schneller sind als der vollständige Worterkenner. Im Gegensatz zum Verfahren in [2] wurden allerdings nicht die 13 Sprecherklassenverteilungen für die Warpingfaktoren auf dem unverzerrten Merkmalsraum bestimmt, sondern eine einzige Klassenverteilung auf dem sprechernormierten Merkmalsraum. Das hat den Vorteil, daß damit der Speicherbedarf für die Verteilungsparameter reduziert wird: statt die Parameter für jede Sprecherklasse zu speichern, wird nur noch eine einzige (normierte) Klasse benötigt. Auf die Geschwindigkeit des Schätzprozesses hat diese Änderung jedoch keinen Einfluß, da wir statt ein unverzerrtes Sprachsignal von 13 Klassifikatoren bearbeiten zu lassen, jetzt 13 verzerrte, sprechernormierte Signale von einem (etwa gleich großen) Klassifikator beurteilen lassen.

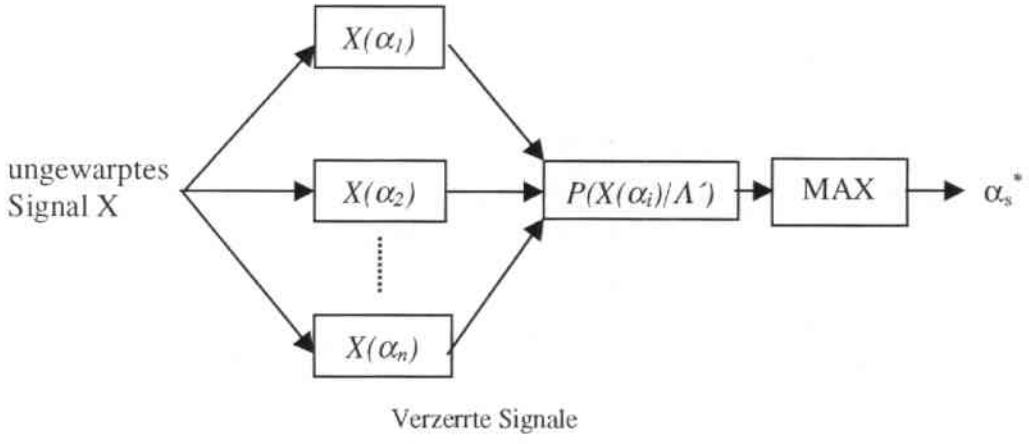

Abbildung 6. Warpingfaktorschätzung nach [1]

Die Wahl des Warpingfaktors erfolgt hierbei gemäß Gleichung (5.3).

(5.3) $\alpha_s^* = argmax_\alpha P(X(\alpha)/\Lambda')$

Das für diesen einzigen Klassifikator benötigte Modell Λ' wird mit sprechernormierten Sprachsignalen $X(\alpha_s)$ trainiert. Dabei werden genau die Warpingfaktoren benutzt, die schon während des Worterkennertrainings für den jeweiligen Sprecher bestimmt wurden.
Mit dieser Vorgehensweise erhalten wir zwar einen Warpingfaktorschätzer, der um ein Vielfaches schneller ist als unser auf dem vollständigen Worterkenner basierter Ausgangsschätzer, allerdings sinkt auch die durch die VTLN erreichte Verbesserung der Worterkennungsrate.

Basissystem (ohne VTLN)	83.9%
VTLN mit Worterkenner (eine Hypothese)	84.8%
VTLN mit 1-Klassenschätzer (keine Hypothese notwendig)	84.1%

Tabelle 4. Erkennungsraten mit dem noch schnelleren 1-Klassenschätzer

Deshalb haben wir nicht nur dieses eine einfache Modell für normierte Sprache getestet, das aus einer einzigen Klasse besteht. In der Hoffnung, einen Warpingfaktorschätzer zu finden, der sowohl schnell ist, als auch eine bessere Schätzung als der 1-Klassenschätzer liefert, haben wir HMMs mit den folgenden Klasseneinteilungen untersucht:

1) Der einfache 1-Klassenschätzer: die einzige Klasse (mit Sprache und Stille) enthält 250 Gaußverteilungen
2) Sprache/Stille: 2 Klassen, eine für Sprache (200 Gaußverteilungen), eine für Stille (50 Gaußverteilungen)
3) stimmhaft/stimmlos: 2 Klassen, eine für stimmhafte Laute (200 Gaußverteilungen), eine für stimmlose Laute und Stille (50 Gaußverteilungen)
4) kontextunabhängiger Phonemerkenner: 68 Phoneme mit je 16 Gaußverteilungen, eine Klasse für Stille

Alle diese Modelle wurden genau so wie schon beim oben beschriebenen einfachen Schätzer auf dem per VTLN sprechernormierten Merkmalsraum trainiert. Sie arbeiteten auf dem gleichen Merkmalsraum wie der Worterkenner (also 28 LDA-Koeffizienten, siehe Kapitel 2) und wurden mit den Viterbi-Pfaden des Worterkenners trainiert. Zum Testen wurde jeder dieser Warpingfaktorschätzer vor den Worterkenner geschaltet und die Erkennungsrate des Worterkenners gemessen. So wie schon in Kapitel 4 erwähnt, gingen in die ML-Scoreberechnung für die Warpingfaktorschätzung bei den HMMs 3 und 4 nur die Wahrscheinlichkeiten der stimmhaften Phoneme ein. Da HMM 2 keine Information über Stimmhaftigkeit liefert, wurden dort nur die als Sprache erkann-

ten Merkmalsvektoren in die ML-Scoreberechnung einbezogen. Der einfache Schätzer (HMM 1) muß alle Merkmalsvektoren zur ML-Scoreberechnung heranziehen.

Basissystem ohne VTLN	83.9%
VTLN basierend auf Worterkenner (nur eine Hypothese)	84.8%
VTLN basierend auf	
1) 1-Klassenschätzer	84.1%
2) Sprache/Stille	84.7%
3) stimmhaft/stimmlos	84.6%
4) kontextunabhängiger Phonemerkenner	84.2%

Tabelle 5. Erkennungsraten mit verschiedenen Warpingfaktorschätzern

	Erzeugung der Hypothese (Viterbi-Suche)	ML-Score-berechnung (13mal)
VTLN basierend auf Worterkenner	68	1.4
VTLN basierend aus Sprache/Stille	0.33	2.4

Tabelle 6. Zeitbedarf verschiedener Warpingfaktorschätzer (in Sekunden Rechenzeit pro Sekunden Sprachdaten)

Wie man Tabelle 5 entnehmen kann, liefert der auf Sprache/Stille basierte Schätzer die besten Worterkennungsraten. Er liegt fast auf dem Niveau des Systems, das den langsamen Worterkenner zur Warpingfaktorschätzung einsetzt. Die schlechte Leistung des auf dem Phonemerkenner basierten Schätzers ist auf die schlechte Erkennungsrate des Phonemerkenners von nur etwa 60% zurückzuführen. Durch seine schlechte Hypothese wird bei der ML-Scoreberechnung für die einzelnen Warpingfaktoren einfach zu häufig die Likelihood für die fehlerhaft bestimmte Phonemklasse in die Rechnung eingebracht. Diese Hypothese haben wir dadurch bestätigt, daß wir in einem Versuch die Transkription nicht durch den Phonemerkenner bestimmen ließen, sondern die richtige Transkription aus der gegebenen Datenbank gelesen haben. Mit diesem 'Schwindel' lieferte der phonemklassenbasierte Schätzer auf der fehlerlosen Transkription noch bessere Ergebnisse als der auf Sprache/Stille basierte Schätzer. An den Ergebnissen fällt auf, daß das stimmhaft/stimmlos HMM keine bessere Leistung bringt, als das Sprache/Stille HMM. Vielleicht hätte ein nicht HMM-basierter Pitchdetektor bessere Ergebnisse geliefert. Die Klärung dieser Frage war aber nicht notwendig, denn die Ergebnisse des Sprache/Stille basierten Schätzers waren sehr zufriedenstellend, so daß wir dieses System als Grundlage für weitere Untersuchungen wählten.

Diese Untersuchungen sollten die Auswirkungen der Wahl verschiedener Parameter auf das Sprache/Stille HMM aufzeigen. Genauer gesagt, die Auswirkun-

gen auf die Erkennungsleistung des von der Warpingfaktorschätzung dieses HMMs abhängigen Worterkenners.

Tabelle 7 zeigt welchen Einfluß die Anzahl der Mixturparameter für die beiden Klassen Sprache und Stille auf die Erkennungsleistung hat. Je mehr Mixturparameter verwendet werden, desto genauer können die Verteilungsfunktionen geschätzt werden, desto langsamer wird allerdings auch der Erkennungsvorgang. Aus der Tabelle läßt sich ablesen, daß 250 Gaußverteilungen eine genügend genaue Approximation der Verteilungen zulassen. Mehr Gaußverteilungen bringen nur eine vernachlässigbare Verbesserung der Erkennungsrate. Aus diesem Grund haben wir die folgenden Experimente mit 250 Gaußverteilungen durchgeführt.

Anzahl der Gaußverteilungen	davon für Sprache	für Stille	
400	300	100	84.7%
250	200	50	84.7%
150	100	50	84.6%

Tabelle 7. Erkennungsraten in Abhängigkeit der Codebuchgröße

Tabelle 8 zeigt, daß Sprache und Stille nicht durch je einen einzigen HMM-Zustand realisiert werden sollten. Durch Verwendung von 9 Zuständen wird eine Art Glättung der resultierenden Sprache/Stille Transkription erreicht, die sich positiv auf die Erkennungsrate auswirkt.

Anzahl HMM-Zustände	Davon für Sprache	für Stille	
2	1	1	84.6%
9	6	3	84.7%
18	12	6	84.7%

Tabelle 8. Erkennungsraten in Abhängigkeit der HMM-Größe

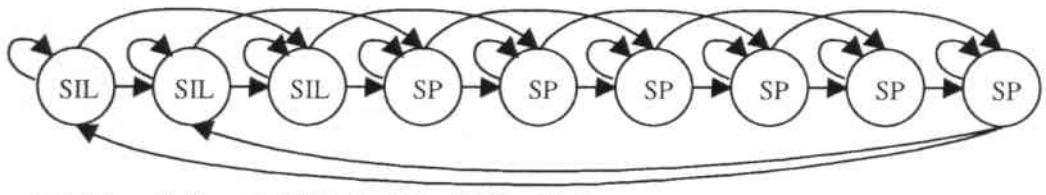

Abbildung 7. Sprache/Stille HMM mit 9 Zuständen

5.2 Schritthaltende Warpingfaktorschätzung

Ziel dieser Studienarbeit war die Untersuchung der ML-VTLN in Bezug auf ihre Verwendbarkeit in schritthaltenden Echtzeiterkennern. Ein Erkenner heißt Echtzeiterkenner, wenn er für die Verarbeitung einer Äußerung nicht mehr Zeit benötigt, als diese Äußerung dauert. Er heißt schritthaltend, wenn die Verzögerung zwischen dem Aussprechen eines Wortes und der Ausgabe der zugehörigen Hypothese klein ist, mit der Erkennung also noch während des Sprechens begonnen wird. Dies hat Einfluß auf Normierungsverfahren, da nicht auf alle Merkmalvektoren der Äußerung zurückgegriffen werden kann. Mit dem Sprache/Stille basierten HMM wurde im letzten Kapitel bereits ein Warpingfaktorschätzer vorgestellt, der einen guten Kompromiß zwischen benötigter Rechenzeit und Erkennungsrate bietet. Ob die Geschwindigkeit des Schätzers für den Einsatz in einem Echtzeiterkenner ausreichend ist, hängt nur noch von der verwendeten Rechenkapazität und der Implementierung der Berechnung der Emissionswahrscheinlichkeiten ab, welche sich zum Beispiel durch Methoden wie BBI [6] beschleunigen läßt.
Ein wichtiger Nachteil der bisherigen Methode ist, daß für die Bestimmung des Warpingfaktors immer erst die gesamte Äußerung von mehreren Sekunden Länge verarbeitet werden mußte. In [3] wird beispielsweise beschrieben, daß die Wahl des Warpingfaktors sogar erst nach durchschnittlich 52 Sekunden erfolgt. In einem schritthaltenden Erkenner ist dieses Verhalten aber im Allgemeinen unerwünscht, da damit die Worthypothese erst nach 52 Sekunden Wartezeit für den Benutzer sichtbar wäre. Unser Ziel war es, die Wartezeit so kurz wie möglich zu halten, ohne die Qualität der Warpingfaktorschätzung zu sehr zu vermindern.
Ein erstes Experiment sollte klären, wie genau der Warpingfaktor geschätzt werden kann, wenn dafür nur ein Teilsegment der Äußerung zur Verfügung steht. Abbildung 8 zeigt die geschätzten Warpingfaktoren für zwei Äußerungen eines Sprechers. Der zum Zeitpunkt t angegebene Warpingfaktor wurde durch das in Abschnitt 5.1 beschriebene Sprache/Stille HMM bestimmt, wobei dem HMM allerdings nur die ersten t Sekunden des Sprachsignals zur Verfügung gestellt wurden.

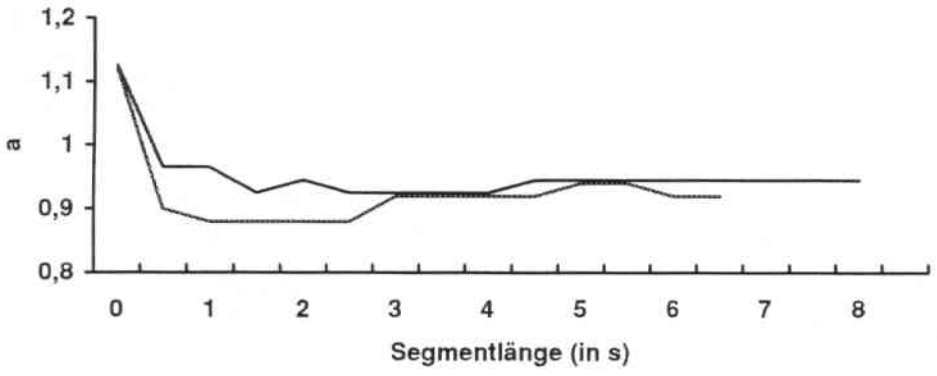

Abbildung 8. Verlauf der Warpingfaktorschätzung bei steigender Segmentlänge

Man sieht, daß der geschätzte Warpingfaktor mit steigender Segmentlänge gegen den sprecherspezifischen Warpingfaktor (im Beispiel bei etwa 0.93) konvergiert. Das heißt der Warpingfaktor für den Sprecher kann umso genauer geschätzt werden, je mehr Daten dem Schätzer zur Verfügung gestellt werden.
Dieser Schluß läßt sich auch aus Abbildung 9 ziehen: diese Abbildung zeigt die durchschnittliche Abweichung des Warpingfaktors in Abhängigkeit der zum Schätzen benutzten Segmentlänge gemittelt über mehrere hundert Äußerungen aus dem GSST Corpus.

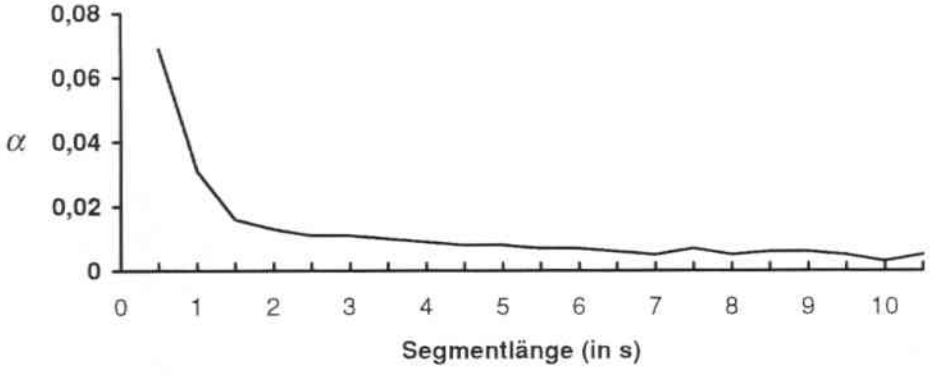

Abbildung 9. Durchschnittliche Abweichung der Warpingfaktorschätzung

Als Ansatz für eine adaptive, schritthaltende Warpingfaktorschätzung wählten wir die Auftrennung des Sprachsignals X in kleinere Sprachsegmente mit fester Länge. Auf jedem dieser gleichlangen Sprachsegmente wurde durch das Sprache/Stille HMM ein Warpingfaktor bestimmt.
Gesucht waren die optimale Sprachsegmentlänge und ein Verfahren, das aus der bis zu einem bestimmten Zeitpunkt bestimmten Warpingfaktorfolge adaptiv einen einzigen Warpingfaktor bestimmt, der letztendlich zum Warpen der Merkmalsvektoren benutzt werden kann. Um schritthaltend zu bleiben, muß das Verfahren möglichst früh eine Warpingfaktorschätzung abgeben, damit der Be-

nutzer, wie oben bereits erwähnt, nicht lange auf die Worterkennerhypothese warten muß. Daß heißt aber auch, daß dem Warpingfaktorschätzer nur wenige Sprachdaten für seine Aufgabe zur Verfügung gestellt werden können, was die Genauigkeit der Schätzung beeinträchtigt.

Tabelle 9 zeigt die Erkennungsergebnisse für Segmentgrößen von 0.5 bis 2 Sekunden, wobei zum Warpen jedes Segments einfach nur derjenige Faktor ausgewählt wurde, der auf genau diesem Segment bestimmt wurde, ohne vergangene Schätzungen zu berücksichtigen.

0.5 s	82.2%
1 s	83.5%
2 s	84.3%

Tabelle 9. Erkennungsraten in Abhängigkeit der Segmentlänge

Die Ergebnisse zeigen, daß für Segmentgrößen unter 2 Sekunden die Erkennungsleistung sogar unter die des Basissystems ohne VTLN (83.9%) fällt. Das ist aber bei Warpingfaktorfolgen mit einer so hohen Varianz zu erwarten, wie das untere Bild verdeutlicht. Die zugehörige Äußerung ist die gleiche wie die der oberen Kurve in Abbildung 2.

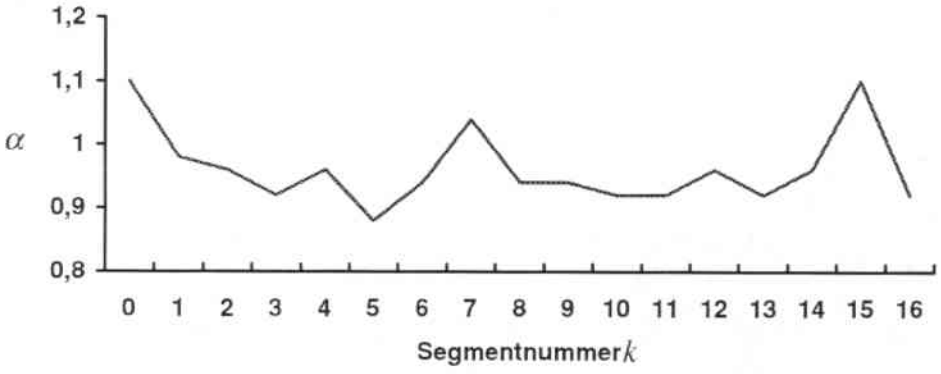

Abbildung 10. Ergebnis der Warpingfaktorschätzung ohne Kontexteinbeziehung

Eine erste Idee zur Verbesserung der Erkennungsleistung war es, eine einfache Medianglättung auf der Warpingfaktorfolge durchzuführen, wodurch aber so gut wie keine Verbesserung erreicht wurde: die Erkennungsleistung blieb unter der des Basissystems. Erfolgreicher war die Methode, nicht die auf den Teilsegmenten berechneten Warpingfaktoren zu kombinieren, sondern die auf jedem Segment für jeden Warpingfaktor berechneten ML-Scores. Dazu wurde die ge-

suchte Emissionswahrscheinlichkeit des die Segmente 0 bis k umfassenden Sprachsignals $X_{[0,k]}$ durch das Sprachmodell Λ approximiert, indem die auf den einzelnen Segmenten berechneten Emissionswahrscheinlichkeiten multipliziert wurden:

$$(5.4)\ P(X_{[0,k]}(\alpha)/\Lambda) \approx \prod_{i=0}^{k} P(X_i(\alpha)/\Lambda')$$

wobei $X_{[0,k]}$ das die Segmente 0 bis k umfassende Sprachsignal ist und X_i das i-te Teilsegment. Mit Hilfe von Approximation (5.4) und Gleichung (5.3) wurde für das k-te Teilsegment X_k ein zugehöriger Warpingfaktor α_k^* gemäß folgender Formel bestimmt:

$$(5.5)\ \alpha_k^* = argmax_\alpha \prod_{i=0}^{k} P(X_i(\alpha)/\Lambda')$$

Diese Formel wurde so implementiert, daß für jeden Warpingfaktor α aus dem Raster die auf allen bisher betrachteten Teilsegmenten 0 bis k berechneten ML-Scores aufaddiert wurden. Segment k wurde dann mit dem Warpingfaktor α_k^* normiert.
Durch diese Methode mit Kontexteinbeziehung (Einbeziehung der Vergangenheit) wird eine ähnlich glatte Warpingfaktorkurve erzeugt, wie in Abbildung 8 dargestellt. Abbildung 11 vergleicht die durch die verschiedenen Methoden gewonnenen Warpingfaktoren.

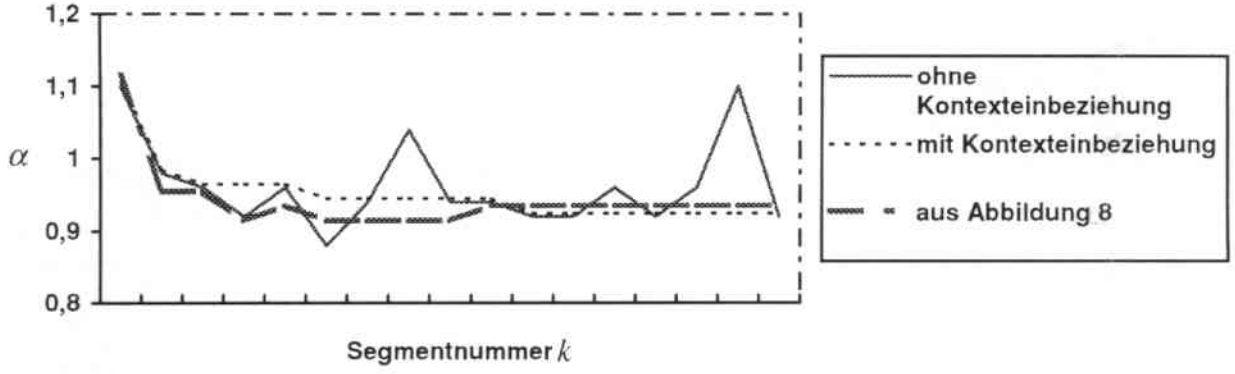

Abbildung 11. Vergleich der Warpingfaktorschätzmethoden

Mit dieser Vorgehensweise ergibt sich allerdings eine notwendige Vorausschau von der Größe eines Segmentes, das heißt der Erkennungsprozeß für Segment i kann erst beginnen, wenn das komplette Signal einschließlich X_i vorhanden ist. In einem weiteren Versuch zeigte sich, daß es ausreicht, diese Vorausschau nur für die Schätzung des Warpingfaktors auf dem allerersten Sprachsegment beizubehalten, die Erkennungsrate wird dadurch um weniger als 0.1 Prozentpunkte beeinträchtigt. Daß heißt, der auf dem ersten 2 Sekunden langen Segment bestimmte Warpingfaktor wird für die ersten beiden Segmente benutzt, der auf

dem zweiten Segment bestimmte Faktor für Segment 3 und so weiter. Mit dieser Vorgehensweise konnte die für den Benutzer unangenehme VTLN-bedingte Verzögerung des Erkennungsvorganges auf die ersten 2 Sekunden Sprache beschränkt werden.
Tabelle 10 vergleicht noch einmal die Ergebnisse der letztgenannten Experimente (die Segmentlänge ist konstant 2 s).

1) ohne Einbeziehung vergangener Messungen, 2s Vorausschau	84.3%
2) mit Einbeziehung vergangener Messungen, 2s Vorausschau	84.5%
3) wie (2), aber Vorausschau nur auf den ersten 2s	84.5%

Tabelle 10. Erkennungsraten in Abhängigkeit der Warpingfaktorselektion

Eine kurze Zusammenstellung der von allen vorgestellten Verfahren gelieferten Ergebnisse kann Tabelle 11 entnommen werden.

Worterkenner basierte VTLN (13 Hypothesen)	85.0%
Worterkenner basierte VTLN (nur eine Hypothese)	84.8%
Sprache/Stille VTLN	84.7%
Sprache/Stille VTLN mit Vorausschau nur auf den ersten 2s	84.5%
Basissystem (ohne VTLN)	83.9%

Tabelle 11. Zusammenfassung der Experimente

In dieser Arbeit wurde die Erkennung nur auf Einzelsätzen von unbekannten Sprechern durchgeführt, das heißt auf jedem Satz waren zwei Sekunden Vorausschau nötig. Es wäre allerdings vorteilhaft zu wissen, daß mehrere aufeinanderfolgende Sätze von ein und demselben Sprecher stammen. Dann könnte man die Vorausschau auf die ersten zwei Sekunden des ersten Satzes von diesem Sprechers beschränken. Außerdem würde sich die Erkennungsrate im Allgemeinen verbessern, weil der Erkenner bei jedem folgenden Satz auf die früher geschätzten Werte für die Warpingfaktorschätzung zurückgreifen kann.
Am Schluß noch ein Wort zu einem Problem, das im Rahmen dieser Studienarbeit nicht angegangen wurde, nämlich der Wechsel von Sprechern während der Erkennung. Wenn der Wechsel dem Erkenner angezeigt würde, könnte die Schätzung des Warpingfaktors einfach neu gestartet werden, indem die Scoreakkumulatoren für jeden Warpingfaktor aus dem Raster wieder mit 0 initialisiert werden, wozu allerdings wieder eine 2 Sekunden dauernde Verzögerung notwendig wären. Falls der Sprecherwechsel aber nicht explizit angezeigt werden könnte, wäre ein zusätzlicher Detektor für dieses Ereignis nötig, was bereits in [3] vorgeschlagen wurde.

A Anhang

A.1 Das Basissystem und der GSST

Die Vorverarbeitung des Basissystems P1 berechnet 13-dimensionale Melscale Cepstrum-Merkmale, die um ihre ersten und zweiten Ableitungen und einen Energiewert ergänzt einen 40-dimensionalen Merkmalsvektor ergeben. Dieser wird einer LDA unterzogen und auf 28 Dimensionen reduziert. Das Wörterbuch umfaßt etwa 6000 Wörter, 68 Phoneme werden unterschieden. Detailliertere Angaben zum Basissystem finden sich in [7].

Verbmobil ist die Bezeichnung für ein Langzeitforschungsprojekt mit der Zielsetzung der Entwicklung eines tragbaren maschinellen Sprachübersetzers, der es etwa Geschäftsleuten aus verschiedenen Sprachräumen ermöglichen soll, in ihrer jeweils eigenen Sprache miteinander zu kommunizieren. Im Rahmen dieses Projektes wurde eine Datenbasis mit etwa 32 Stunden transkribierter spontaner deutscher Sprache als Trainingsmaterial erstellt. Um eine repräsentative Mischung verschiedener deutscher Dialekte zu erlangen, wurden die Sprachdaten an vier verschiedenen Orten innerhalb Deutschlands gesammelt. Obwohl die Domäne eingeschränkt ist, wurden keinerlei Restriktionen bei der Auswahl der Sprecher oder Sprachstile angewendet. Typische Phänomene spontan gesprochener Äußerungen, wie etwa Hintergrundgeräusche, Stottern, unvollständige oder grammatikalisch falsche Sätze sind daher keine Seltenheit. Was die Sprachmodellierung angeht, ist der Verbmobil-Korpus eher klein. Die Äußerungen umfassen etwa 300000 Wörter, das Vokabular ist etwa 6000 Wörter groß.

Der German Spontaneous Scheduling Task (GSST) ist ein auf der Verbmobil-Datenbasis aufbauender Benchmark mit einer Trainingsmenge von 14009 Äußerungen und einer Testmenge von 343 Äußerungen.

Weitere Angaben zum Verbmobil-Korpus und zum GSST findet man etwa in [7]. Dort sind auch die Resultate der GSST-Evaluationen des Janus RTK und anderer Systeme aus den Jahren 1995 und 1996 aufgeführt.

Literatur

[1] Zhan,Westphal
„Speaker Normalization Based On Frequency Warping"
ICASSP´97 S.1039 ff.

[2] Lee,Rose
„Speaker Normalization Using Efficient Frequency Warping Procedures"
ICASSP´96 S.353 ff.

[3] Wegmann,Allaster,Orloff,Peskin
„Speaker Normalization on Conversational Telephone Speech"
ICASSP´96 S.339 ff.

[4] Eide,Gish
„A Parametric Approach to Vocal Tract Length Normalization"
ICASSP´96 S.346 ff.

[5] „Spracherkennung bei Daimler-Benz"
Verbmobil Akustik Workshop, Herrenberg 9. und 10. Oktober 1997.

[6] Fritsch,Rogina
The Bucket-Box-Intersection Algorithm for fast approximative Evaluation of Diagonal Mixture Gaussians.
ICASSP´96, Atlanta.

[7] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, Martin Westphal.
„The Karlsruhe Verbmobil Speech Recogition Engine"
In IEEE Conference on Acoustics, Speech and Signal Processing, 1997.