Sprachsynthese mit spezieller Prosodie

Reto Pieren

Semesterarbeit SA-2007-08

Wintersemester 2006/07

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: M. Gerber, Dr. B. Pfister, H. Romsdorfer

Verantwortlicher: Prof. Dr. L. Thiele

Inhaltsverzeichnis

Abstract

In diesem Bericht wird ein Verfahren vorgestellt, mit dem die Prosodie eines zu synthetisierenden Sprachsignals benutzerfreundlich vorgegeben und vom TTS-System übernommen werden kann. Der gewünschte Text wird dem System mit der gewünschten Prosodie vorgesprochen, worauf dieses die Prosodie des vorgesprochenen Textes ermittelt und bei der Synthese einsetzt. Zur Ermittlung der Prosodie wird eine Erweiterung des *Dynamic Time Warpings* verwendet. Diese Erweiterung arbeitet nicht mit globalen, sondern mit adaptiven Kontinuitätsbedingungen und wird deshalb *Adaptive Dynamic Time Warping* genannt.

1 Einleitung

Auf dem Markt existieren bereits Sprachsynthesesysteme, die aus einem gegebenen Text ein Sprachsignal erzeugen. Diese so genannten Text-To-Speech-Systeme oder TTS-Systeme werden üblicherweise für das neutrale Vorlesen eines Textes ausgelegt. In einigen Anwendungen, wie zum Beispiel einem automatischen Auskunftssystem, kann aber für einige Sätze auch eine spezielle Sprechweise erwünscht sein. Ein Begrüssungstext zum Beispiel sollte eher freundlich und einladend als neutral gesprochen werden. Auch ein Firmenname sollte vielleicht speziell betont werden. In solchen Fällen sollte die Sprechweise, auch Prosodie genannt, vom Anwender definiert werden können.

Unter Prosodie werden in der Sprachverarbeitung meist drei Komponenten, prosodische Grössen genannt, verstanden. Es sind dies die Dauer der Laute, die Grundfrequenz und die Intensität. In natürlicher Sprache werden diese drei Komponenten vom Sprecher individuell gesteuert und erzeugen so eine bestimmte Sprechweise. In der Dauer der Laute oder Halblaute ist auch die Information über Sprechrhythmus und Sprechtempo enthalten.

Manuelles Verändern von Parametern setzt grosses Vorwissen im Bereich Linguistik, Akustik, Signalverarbeitung usw. voraus. Deshalb bietet sich diese Methode der Modifikation nicht an. Auch die Idee, für die speziell auszusprechenden Sätze anstelle der synthetischen Stimme eine natürliche Sprachaufnahme zu verwenden, muss verworfen werden, da die Stimme des eingefügten, natürlich gesprochenen Ausschnitts nicht mit jener der Sprachsynthese zusammenpasst. Eine benutzerfreundliche Vorgabe der gewünschten Prosodie an das TTS-System ist also erstrebenswert.

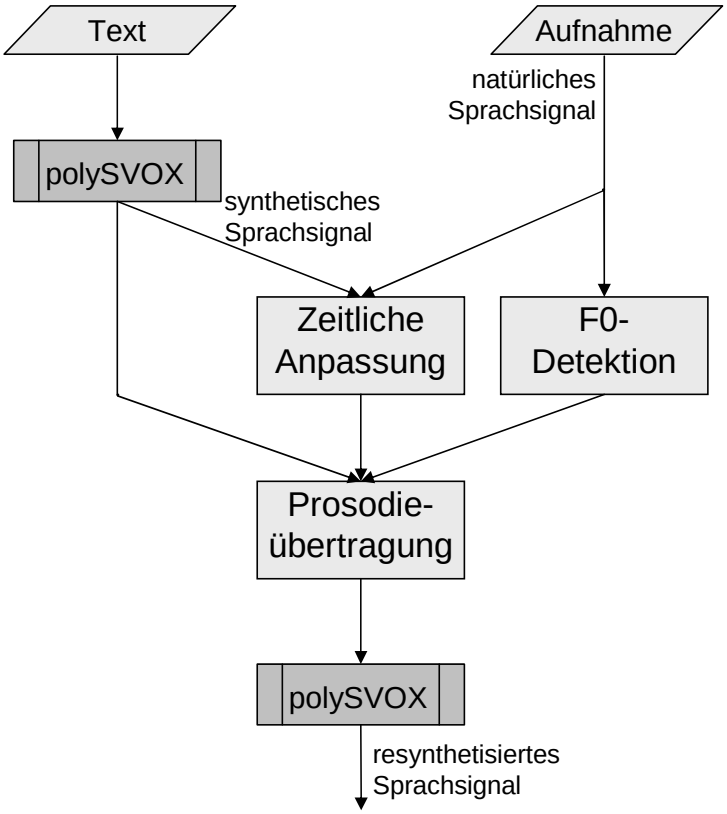
Am einfachsten kann die Prosodie durch Vorsprechen vorgegeben werden. Genau diese Idee wurde in der vorliegenden Arbeit aufgegriffen. Der zu modifizierende Satz wird dem System mit der gewünschten Prosodie vorgesprochen. Diese Prosodie soll auf das Sprachsignal der synthetischen Stimme übertragen werden. Die Übertragung soll so sein, dass der modifizierte Satz zu den anderen, vom System neutral vorgelesenen Sätzen passt. Unter anderem soll die Tonlage gleich bleiben. Zudem soll der Satz korrekt ausgesprochen werden, auch wenn im vorgesprochenen Satz einige Laute verschluckt wurden.

Zusätzlich zu diesem schriftlich vorliegenden Bericht wurde eine Audio-CD erstellt. Diese enthält Beispiele von den im Bericht behandelten Sprachsignalen. Mit [CD:T01] wird auf Track 01 dieser Audio-CD verwiesen.

2 Beschreibung des Systems

2.1 Blockdiagramm

In Figur 1 ist ein vereinfachtes Blockdiagramm des Systems dargestellt. Die Eingänge des Systems sind einerseits der zu synthetisierende *Text*, andererseits wird dem System eine *Aufnahme* übergeben, die ein Sprachsignal desselben Textes enthält. Das vorgesprochene Sprachsignal enthält die vom Anwender gewünschte Prosodie und wird im weiteren Verlauf als *natürliches Sprachsignal* bezeichet. Der Ausgang des Systems ist ein synthetisches Sprachsignal des vorgegebenen Textes mit der vom Anwender vorgesprochenen Prosodie. Dieses Signal wird im weiteren Bericht als *resynthetisiertes Sprachsignal* bezeichnet.



Figur 1: *Vereinfachtes Blockdiagramm des Systems*

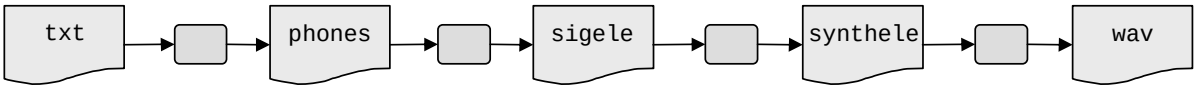
Als erster Schritt wird aus dem vorgegebenen Text ein *synthetisches Sprachsignal* erzeugt. Dazu wird das Sprachsynthesesystem polySVOX eingesetzt, welches in Kapitel 2.2 beschrieben wird. Danach werden die Merkmale des natürlichen und des synthetischen Sprachsignals extrahiert. Mit Hilfe der extrahierten Merkmale werden die Sprachsignale miteinander verglichen und zeitlich aneinander angepasst. Dieser Prozess wird in Figur 1 mit dem Block *Zeitliche Anpassung* repräsentiert und in Kapitel 4 beschrieben. Die Detektion des Grundfrequenzverlaufs wird mit dem Block *F0-Detektion* symbolisiert und in Kapitel 5.3 beschrieben. Dann wird die eigentliche *Prosodieübertragung* durchgeführt, indem die synthetische Prosodieinformation modifiziert wird. Nach der Prosodieübertragung wird erneut eine Sprachsynthese mit

polySVOX durchgeführt. Das neu synthetisierte Signal entspricht dem oben erwähnten resynthetisierten Sprachsignal und enthält die vom Anwender gewünschte Prosodie.

2.2 TTS-System polySVOX

In dieser Arbeit wurde das am Institut für Technische Informatik der ETH Zürich entwickelte Text-To-Speech-System polySVOX eingesetzt. Dieses System kann aus einem schriftlich vorliegenden, deutschen Text ein synthetisches Sprachsignal erzeugen. Ein ausführlicher Beschrieb zum Funktionsprinzip von polySVOX ist in [Tra95] zu finden. Hinweise zur Anwendung der Programm-Umgebung finden sich in [Rom06].

In den folgenden Unterkapiteln wird kurz auf die verschiedenen Ein- und Ausgabedateien von polySVOX eingegangen. Ein Datenflussdiagramm der Ein- und Ausgabedateien ist in Figur 2 zu sehen, denn die verschiedenen Dateiformate spielen im Zusammenhang mit dieser Arbeit eine wichtige Rolle.



Figur 2: *Datenflussdiagramm von polySVOX*

2.2.1 txt-Datei

Als primäre Eingabe ist eine `txt`-Datei vorgesehen. Diese enthält den zu synthetisierenden Text.

2.2.2 phones-Datei

Die `phones`-Datei enthält die Phoneme des zu synthetisierenden Signals. Sie kann als Eingabe- wie auch als Ausgabeformat gewählt werden. Nebst der Bezeichnung der Phoneme ist für jedes Phonem die Dauer in Millisekunden und der Grundfrequenzverlauf in Hertz mit 5 Stützwerten angegeben. Ein Ausschnitt einer `phones`-Datei ist in Anhang B zu finden.

2.2.3 sigele-Datei

In der `sigele`-Datei sind die Halbdiphon-Elemente enthalten. Diese Elemente werden von polySVOX aneinandergehängt, um das gewünschte Sprachsignal zu erzeugen. Für jedes Halbdiphon-Element sind unter anderem die Dauer in Millisekunden und 5 Stützwerte des Grundfrequenzverlaufs angegeben. In Anhang B ist ein Ausschnitt einer `sigele`-Datei zu finden.

2.2.4 synthele-Datei

Die `synthele`-Datei enthält genau die gleichen Halbdiphon-Elemente wie die `sigele`-Datei, jedoch wird diese Datei erst *nach* der Synthese erzeugt und enthält somit die wirklich realisierten Dauerwerte der Elemente. Die in der `sigele`-Datei angegebenen Werte werden von polySVOX nicht exakt realisiert. Der Grund liegt einerseits darin, dass der verwendete PSOLA-Algorithmus stets ganze Signalperioden aneinanderreiht. Andererseits können gewisse Halbdiphone vom Algorithmus nicht unbegrenzt gestreckt oder gestaucht werden. Diese wirklich erzeugten Halbdiphon-Dauern variieren für diese Anwendung zu stark von den in der `sigele`-Datei angegebenen Werten. Deshalb wird für die Prosodiedetektion das `synthele`-Format verwendet. Dieses Dateiformat ist als reines Ausgabeformat von polySVOX zu verstehen.

2.2.5 wav-Datei

Die schlussendliche Ausgabe von polySVOX ist eine `wav`-Datei. Dieses Audiosignal enthält den von der synthetischen Stimme gesprochenen Text. Zwei `wav`-Dateien sind in Figur 13 abgebildet.

3 Merkmalsextraktion und Distanzmass

3.1 Einleitung

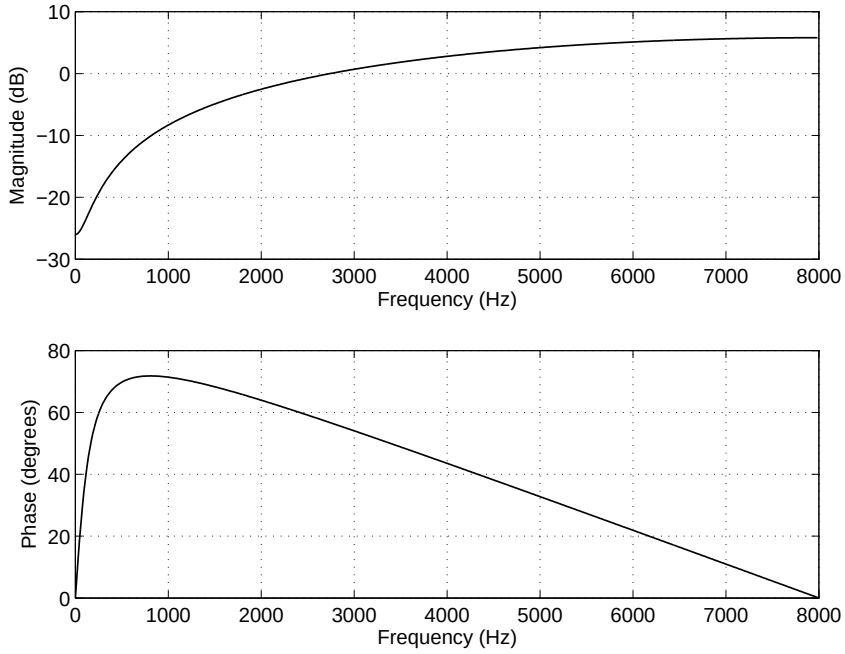
Damit das natürliche und das synthetische Sprachsignal zeitlich aneinander angepasst werden können, wurden geeignete Merkmale aus den Sprachsignalen extrahiert. Anschliessend musste ein Distanzmass definiert werden, damit die extrahierten Merkmalssequenzen miteinander verglichen werden konnten.

3.2 Präemphase-Filter

Auf das synthetische und das natürliche Sprachsignal wurde vor der Merkmalsberechnung das in [PB05] beschriebene Präemphase-Filter angewendet. Die Übertragungsfunktion lautet:

$$H_p(z) = 1 - \beta z^{-1} \tag{3.1}$$

Für β wurde der Wert 0.95 gewählt. Figur 3 zeigt das Bode-Diagramm dieses Filters.



Figur 3: *Bode-Diagramm des Präemphase-Filter $H_p(z)$*

3.3 Merkmalsberechnung

Die gefilterten Signale wurden mit einem Hamming-Fenster der Länge 25ms und einem Shift von 5ms gefenstert. Für jedes so erhaltene Frame wurden dann die Merkmale berechnet. Auf Grund des Shifts von 5ms resultierte eine maximale zeitliche Auflösung von 5ms.

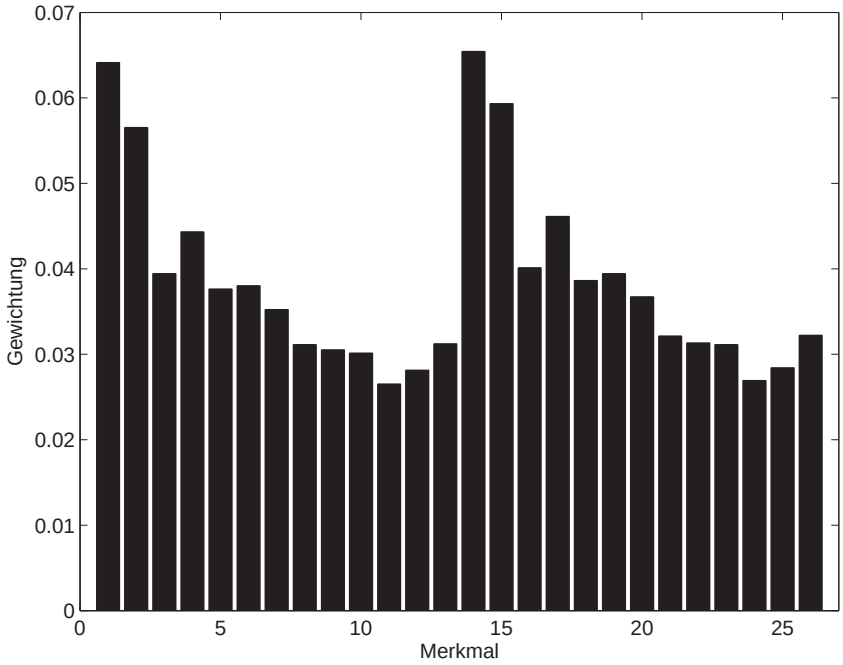
Als Merkmale wurden primär die auch in der Spracherkennung eingesetzten cepstralen Koeffizienten (MFCCs) verwendet. Die theoretischen Abhandlungen und die Berechnungsformeln sind in [PB05] zu finden. Es wurden 34 Mel-Filterbänke eingesetzt und insgesamt 13 cepstrale Koeffizienten inklusive $c(0)$ berechnet.

Weitere Merkmale wurden aus dem Delta-Cepstrum gewonnen, welches eine Approximation der zeitlichen Ableitung der cepstralen Koeffizienten darstellt. Für die Approximation der Ableitung wurden für jedes Frame die zwei Frames davor sowie die zwei Frames danach verwendet.

So ergaben sich pro Frame 26 Merkmale (13 cepstrale Koeffizienten und 13 Approximationen der zeitlichen Ableitung der cepstralen Koeffizienten). Diese 26 Merkmale bildeten pro Frame einen Merkmalsvektor der Länge 26.

3.4 Normalisierung und Gewichtung

Da sich der Übertragungskanal des natürlichen und des synthetischen Sprachsignals sehr unterscheiden, wurden die 26 Merkmale im Sinne einer Kanalkompensation entlang der Merkmale mittelwertbereinigt (engl. Cepstral Mean Subtraction). Zusätzlich wurden die Merkmale auf Varianz 1 gebracht. Diese normalisierten Merkmale wurden dann mit einer empirisch gefundenen Gewichtung versehen. Die verwendete Gewichtung ist in Figur 4 abgebildet.



Figur 4: *Gewichtung der Merkmale*

Die Gewichtung der Merkmale wurde wie folgt gefunden: Die Prosodieübertragung wurde zuerst ausschliesslich mit einer Mittelwertbefreiung der Merkmale durchgeführt. Für einige Signale war die zeitlichen Anpassung so bereits genug gut. Diese Signale konnten als Trainingsdaten für die Bestimmung der Gewichtung verwendet werden. Aus den Trainingsdaten wurde

für jedes Merkmal i dessen mittlerer quadratischer Fehler MSE_i berechnet. Das i-te Gewicht ist eine skalierte Version von $\frac{1}{\sqrt{MSE_i}}$. Die Gewichte wurden so skaliert, dass deren Summe gleich 1 ist. Es wurden also jene Komponenten des Merkmalsvektors, welche viel zum Fehler beitragen, abgeschwächt. Auffällig ist, dass die Merkmale aus dem Delta-Cepstrum ($i = 14, \ldots, 26$) ungefähr dieselben Gewichte wie die Merkmale aus dem Cepstrum ($i = 1, \ldots, 13$) aufwiesen. Mit dieser Gewichtung konnten auch Signale, für welche die zeitliche Anpassung schwierig war und die nicht zur Berechnung der Gewichte verwendet wurden, besser behandelt werden.

3.5 Distanzmass

Für die Berechnung der Distanzmatrix, welche in Kapitel 4.3.1 beschrieben wird, wird ein geeignetes Distanzmass benötigt. In dieser Arbeit wurde als Distanzmass die euklidische Distanz verwendet.

Wie in der Aufgabenstellung in Anhang A erwähnt, kann auch ein komplexeres Distanzmass verwendet werden. Hier bietet sich insbesondere die Verwendung eines neuronalen Netzes an. Mit neuronalen Netzen wurden unter anderem in [GM05] gute Ergebnisse erzielt. Die Schwierigkeit beim Einsatz von neuronalen Netzen ist das Training des Netzes, denn dafür werden viele Trainingsdaten benötigt. Da solche Trainingsdaten nicht verfügbar waren, musste im Zusammenhang mit dieser Arbeit leider auf den Einsatz von neuronalen Netzen zur Distanzberechnung verzichtet werden. Es wird jedoch vermutet, dass mit einem gut trainierten neuronalen Netz die Qualität der zeitlichen Anpassung gesteigert werden könnte.

4 Zeitliche Anpassung der Signale

4.1 Einleitung

Damit die Lautdauern des natürlichen Sprachsignals ermittelt werden können, werden das natürliche und das synthetische Sprachsignal zeitlich aneinander angepasst.

Das natürliche und das synthetische Sprachsignal lassen sich nicht durch eine globale lineare Streckung anpassen. Beim Sprechen werden die Dauern der einzelnen Laute je nach Sprechweise ganz individuell gestreckt oder gekürzt. Beispielsweise werden bei tiefem Sprechtempo Vokale wie [a] oder [y] deutlich stärker zeitlich gestreckt als Plosive wie [p]. Die statische zeitliche Anpassung ist hier keinesfalls anwendbar, es bedarf einer dynamischen zeitlichen Anpassung. Ein Optimierungsverfahren, das die dynamische zeitliche Anpassung zweier Merkmalssequenzen ermittelt, ist *Dynamic Time Warping*. In Kapitel 4.3 wird dieses Verfahren, das in dieser Arbeit angewendet wurde, genauer betrachtet. Vor dem Dynamic Time Warping wurde eine *Utterance Detection* durchgeführt. Dieses Verfahren wird im folgenden Kapitel beschrieben.

4.2 Utterance Detection

Aus den Sprachsignalen wurden die Bereiche, welche Sprache enthalten extrahiert und die Bereiche vor und nach der eigentlichen Äusserung weggeschnitten. Die Äusserung des Sprechers soll so von vorangehenden und gefolgten Störgeräuschen getrennt werden, damit die anschliessende Analyse überhaupt möglich ist. Die Äusserungsdetektion (engl. Utterance detection) wurde offline durchgeführt, für die Ermittlung stand also das ganze Signal zur Verfügung.

Beim synthetisierten Signal ist die Äusserungsdetektion einfach. Aus der `synthele`-Datei kann der Anfangs- und Endzeitpunkt der Äusserung grob ausgelesen werden. Für die exakte Ermittlung genügt dann eine Detektion mittels Intensitäts-Threshold.

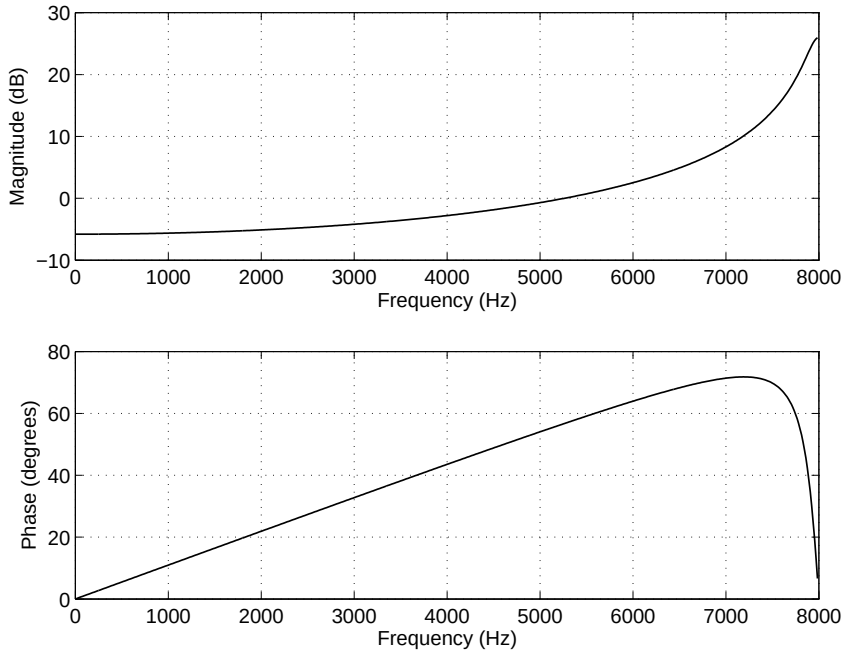
Um die Äusserung im natürlichen Sprachsignal zu detektieren, wurde zuerst ein Vorfilter $H_u(z)$ angewendet. Dieses wird in Kapitel 4.2.1 beschrieben. Anschliessend wurde mit dem gefilterten Signal eine Anfangs- und Endpunktdetektion durchgeführt. Dieses Verfahren wird in Kapitel 4.2.2 detailliert beschrieben.

4.2.1 Filter $H_u(z)$

Um intensitätsarme Frikative wie [s] und [f] am Anfang und am Ende der Äusserung zu verstärken, wurde vor der eigentlichen Utterance Detection ein Filter mit folgender Übertragungsfunktion angewendet:

$$H_u(z) = \frac{1}{1 + \alpha z^{-1}} \tag{4.1}$$

Es wurde $\alpha = 0.95$ gewählt. Mit diesem Filter werden die hohen Frequenzen angehoben. Ein Bode-Diagramm ist in Figur 5 abgebildet.



Figur 5: *Bode-Diagramm des Filters* $H_u(z)$

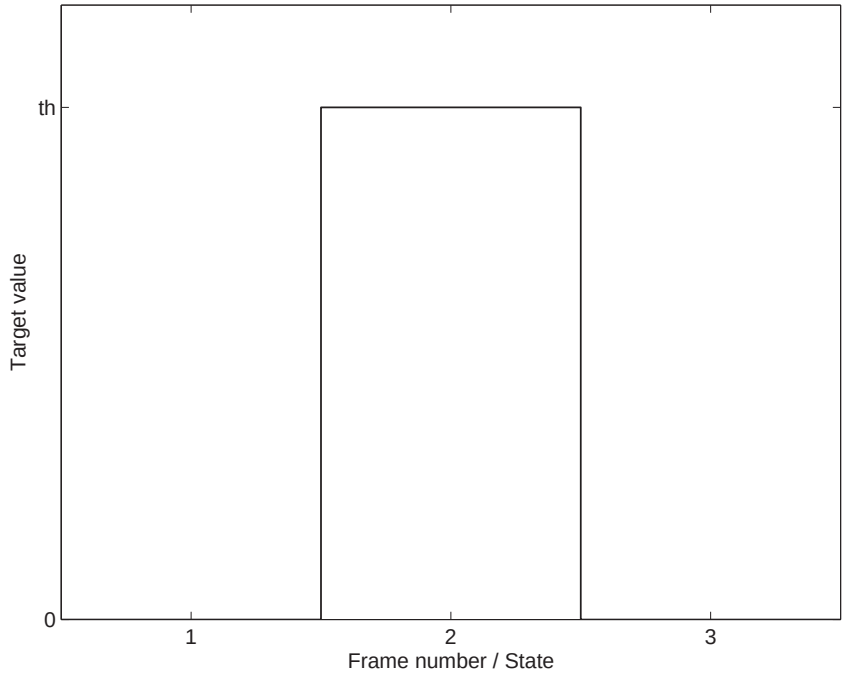
4.2.2 Anfangs- und Endpunktdetektion

Nach dem Filter $H_u(z)$ wurde für die Erkennung der Äusserung ein Verfahren zur Bestimmung des Anfangs- und Endzeitpunktes der Äusserung angewendet. Dieses Verfahren ist dem in Kapitel 4.3 beschriebenen Dynamic Time Warping sehr ähnlich.

Aus dem Sprachsignal wurde mittels Fensterung (Fensterlänge = 6ms, Shift = 3ms) und anschliessender Intensitätsberechnung der grobe Intensitätsverlauf des Signals bestimmt. Dieser Intensitätsverlauf sollte dann einer Targetfunktion, wie in Figur 6 gezeigt, zeitlich angepasst werden. Diese Targetfunktion besteht aus drei Abschnitten, welche die drei zeitlichen Zustände (1: vor Äusserung [Intensität = 0], 2: Äusserung [Intensität = th], 3: nach Äusserung [Intensität = 0]) des natürlichen Sprachsignals darstellen. Mit einem Optimierungsverfahren wurde das Targetsignal dem natürlichen Signal angepasst. Die Zeit, in der das Targetsignal im Zustand 2 verharrt, entspricht der Zeit der Äusserung. Die zwei Zustandsübergänge (von Zustand 1 nach 2 und von Zustand 2 nach 3) in der Targetfunktion wurden dann auf das Sprachsignal abgebildet. Diese zwei Stellen im Sprachsignal wurden als Äusserungsbeginn, respektive als Äusserungsende interpretiert.

4.3 Dynamic Time Warping

Dynamic Time Warping ist ein Verfahren, das mit den Werkzeugen der Dynamischen Programmierung ein Optimierungsproblem löst. Mit Dynamic Time Warping können zwei Merkmalssequenzen zeitlich aneinander angepasst werden. Eine typische Anwendung aus der Sprachverarbeitung ist der Einsatz in einem Spracherkenner zur Erkennung isolierter Wörter. Eine ausführliche theoretische Abhandlung über Dynamic Time Warping ist in [PB05] zu finden. In der



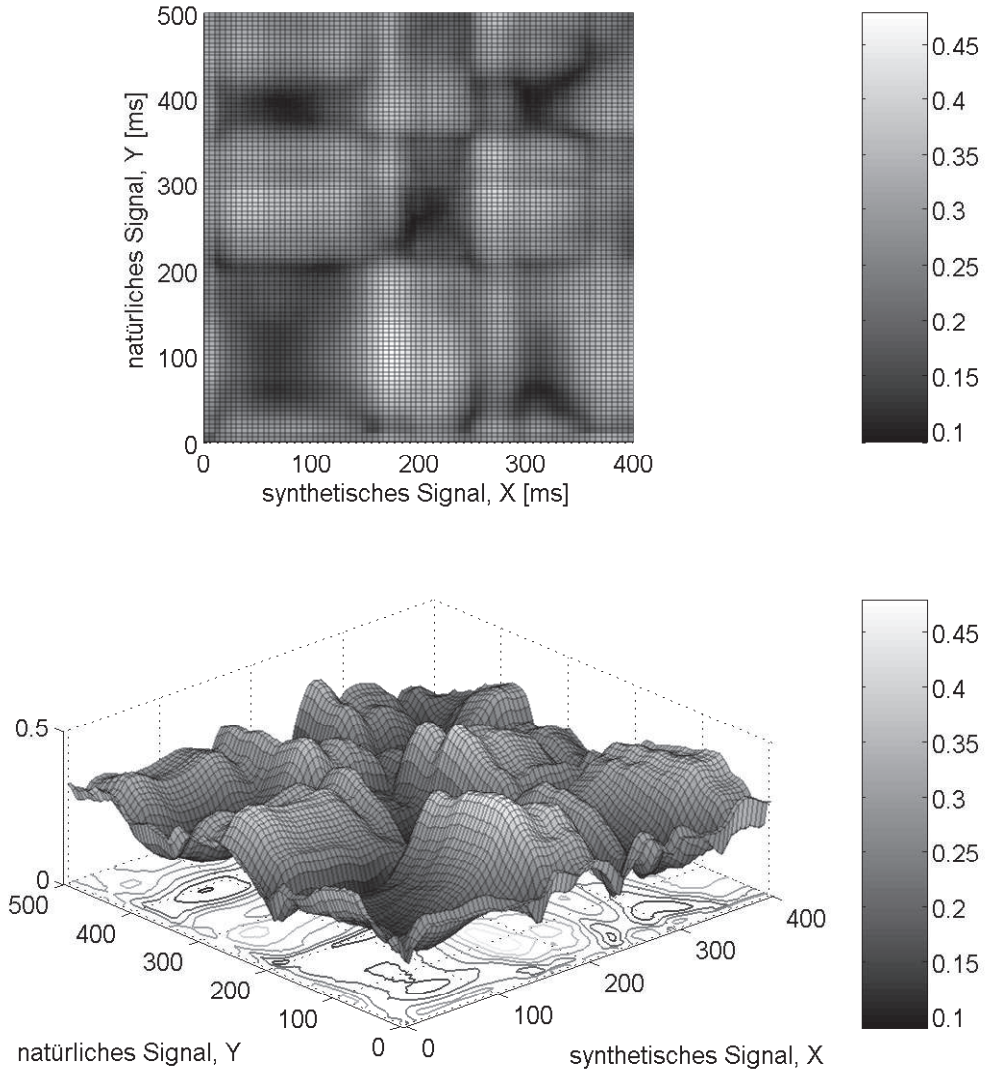
Figur 6: *Target-Funktion der Utterance Detection*

Literatur, wie zum Beispiel in [KE01], ist auch eine Variante des Dynamic Time Warpings unter dem Namen Derivative Dynamic Time Warping zu finden. Diese Variante verwendet zusätzlich zu den eigentlichen Merkmalen deren approximierte Ableitungen. Letztere wurden auch in dieser Arbeit verwendet. Da es sich aber effektiv nur um eine Merkmals-Erweiterung handelt, wird dieser Begriff im Weiteren nicht verwendet.

4.3.1 Distanzmatrix

Vor der Anwendung von Dynamic Time Warping wird aus den beiden Merkmalssequenzen Y und X die sogenannte Distanzmatrix bestimmt. Jedes Element d_{ij} dieser Matrix enthält die Distanz der Merkmalsvektoren Y_i und X_j. Die Art des Distanzmasses wird in Kapitel 3.5 besprochen. Die Distanzmatrix kann als dreidimensionale Fehlerlandschaft interpretiert werden.

In Kapitel 3.3 wurde die Gewinnung der Merkmalsvektoren der beiden anzupassenden Signale beschrieben. Die Merkmalsvektoren des synthetischen Sprachsignals werden zur Merkmalssequenz X, diejenigen des natürlichen Sprachsignals zur Merkmalssequenz Y zusammengefasst. Diese beiden Sequenzen sollen aneinander angepasst werden. Jedes Element d_{ij} der Distanzmatrix enthält also die Distanz von Merkmalsvektor von Frame i des synthetischen Signals und Merkmalsvektor von Frame j des natürlichen Sprachsignals. Je kleiner das Distanzmatrixelement d_{ij} ist, desto ähnlicher sollten die gesprochenen Inhalt der beiden dazugehörigen Frames aus den beiden Sprachsignalen idealerweise sein. In Figur 7 ist ein Ausschnitt einer Distanzmatrix zu sehen. Der untere Teil der Figur zeigt die dreidimensionale Fehlerlandschaft. Der obere Teil der Figur zeigt den gleichen Ausschnitt der Distanzmatrix und kann als Projektion der Fehlerlandschaft auf die XY-Ebene verstanden werden.

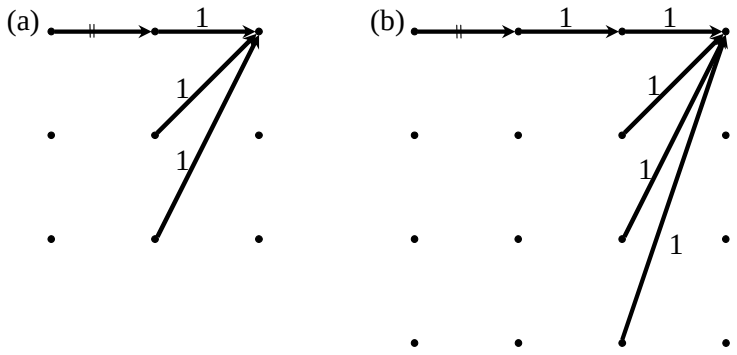


Figur 7: *Zwei Darstellungen der gleichen Distanzmatrix*

4.3.2 Kontinuitätsbedingungen

Innerhalb der vorangehend erläuterten Distanzmatrix soll der optimale Pfad gefunden werden. Dieser Pfad wird auch Warping-Kurve genannt und stellt eine Zuordnung der Merkmalsvektoren aus den beiden Merkmalssequenzen Y und X dar. Der Pfad innerhalb der Distanzmatrix muss gewisse Randbedingungen, die sogenannten Kontinuitätsbedingungen, erfüllen.

Die *Anfangs- und Endpunktbedingung* besagt, dass der Pfad an der Stelle $(1, 1)$ in der Distanzmatrix beginnen und beim Element (Ly, Lx) enden soll. Ly und Lx entsprechen dabei der Anzahl Frames im synthetischen, beziehungsweise im natürlichen Sprachsignal. In der Distanzmatrix aus Figur 7 heisst dies, dass der Pfad in der unteren linken Ecke beginnen und in der oberen rechten Ecke enden muss. Diese Bedingungen bewirkt, dass der Anfangs- und der Endzeitpunkt der Äusserungen aus den zwei Signalen einander zugeordnet werden. Dies bedingt

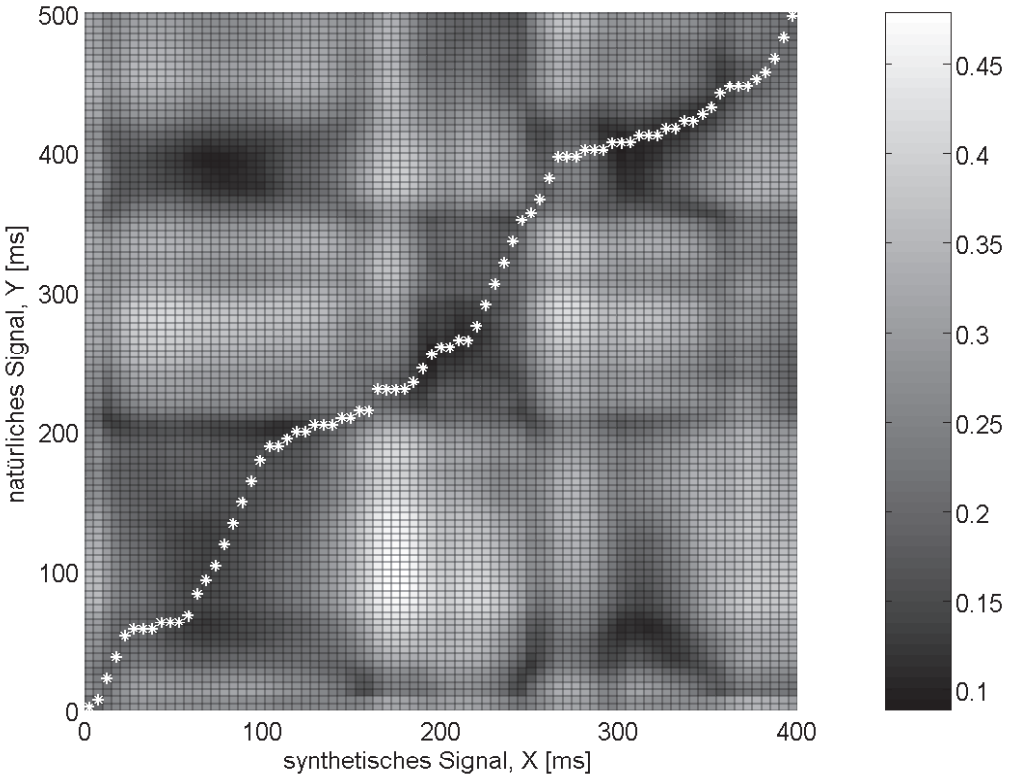


Figur 8: *Zwei Typen von asymmetrischen Pfaderweiterungen (Dynamic Time Warping)*

natürlich, dass die Äusserungen mittels der in Kapitel 4.2 beschriebenen Utterance Detection genug gut ausgeschnitten wurden. Eine Möglichkeit, diesen hohen Qualitätsanspruch an die Utterance Detection etwas zu verringern, wäre die Einführung von sogenannten Delta-Elementen in die Distanzmatrix.

Weitere Bedinungen an den optimalen Pfad sind die erlaubten *Pfaderweiterungen*. Diese werden global gewählt und definieren rekursiv die erlaubten Vorgängerstellen der Stelle (i, j) in der Distanzmatrix. Bei der Bildung des optimalen Pfades besagen die Pfaderweiterungen, welche Erweiterungen zulässig sind. Mit den Pfaderweiterungen wird also auch die maximale und minimale Steigung des Pfades bestimmt. Zudem können die einzelnen Pfaderweiterungen gewichtet werden, womit gewisse Richtungen bevorzugt oder benachteiligt werden können. Aus diesen Gründen haben die Pfaderweiterungen einen sehr grossen Einfluss auf den gefundenen optimalen Pfad und können sehr verschieden gewählt werden. Es war eine grosse Herausforderung, eine für die Anwendung angebrachte Wahl der Pfaderweiterungen zu treffen. Die Wahl erforderte viele Experimente und schliesslich genaue Kenntnis über die verwendeten Daten und die dazugehörigen Distanzmatrizen.

Bei der Anwendung von Dynamic Time Warping haben sich zwei Typen von Pfaderweiterungen als zuverlässig erwiesen. Diese sind in Figur 8 zu sehen. Diese Art von asymmetrischen Kontinuitätsbediungen hat sich auch aus algorithmischen Gründen sehr bewährt. Die Gewichte sämtlicher Pfade sind 1, es wird also grundsätzlich keine Erweiterung der anderen vorgezogen. Der doppelt durchgestrichene Pfad bedeutet, dass diese Erweiterung an dieser Stelle nicht vorkommen darf. In der Pfaderweiterung (a) der Figur 8 dürfen nie zwei horizontale Erweiterungen hintereinander vorkommen. In der Pfaderweiterung (b) hingegen dürfen nie mehr als zwei horizontale Erweiterungen unmittelbar hintereinander vorkommen. Mit diesen Bedingungen wird ein allzu flacher Verlauf des Pfades vermieden und die minimale Steigung, über mehrere Frames betrachtet, erhöht. Die Pfaderweiterung (a) der Figur 8 ermöglicht globale Steigungen im Bereich [1/2, 2]. Pfaderweiterung (b) der Figur 8 ist hingegen liberaler und erlaubt globale Steigungen im Bereich [1/3, 3]. Zweckdienliche Modifikationen dieser Kontinuitätsbedingungen sind in Kapitel 4.4 beschrieben.



Figur 9: *Optimaler Pfad durch die Distanzmatrix aus Figur 7 mit den Pfaderweiterungen (b) aus Figur 8*

4.3.3 Optimaler Pfad

In der Distanzmatrix wird unter Einhaltung der Kontinuitätsbedingungen der optimale Pfad gesucht. Der optimale Pfad ist derjenige Pfad, dessen Summe der durchwanderten Distanzmatrixelemtente d_{ij} am kleinsten ist. Der optimale Pfad stellt also ein globales Minimum dar und liefert unter den gemachten Kontinuitätsbedingungen die beste zeitliche Anpassung der beiden Merkmalssequenzen. Gelöst wird dieses Optimierungsproblem mit Dynamischer Programmierung. In Figur 9 ist der optimale Pfad in die Distanzmatrix aus Figur 7 eingezeichnet.

Der optimale Pfad liefert also eine Zuordnung der Frames des synthetischen Sprachsignals und der Frames des natürlichen Sprachsignals. Wenn das synthetische Sprachsignal als Referenz herangezogen wird, bedeutet ein Bereich mit grosser Steigung des optimalen Pfades eine grosse Stauchung des natürlichen Signals. Umgekehrt bedeutet ein flacher Verlauf des optimalen Pfades eine Streckung des natürlichen Signals. Konkret bedeutet eine Pfadsteigung von 2 eine Signalstauchung um den Faktor 0.5.

Für die meisten in dieser Arbeit verwendeten Daten genügten die Pfaderweiterungen (a) aus Figur 8. Das heisst die Streckungsfaktoren der Dauerwerte bewegen sich im Bereich [1/2, 2]. Für einige extrem betonte Stellen reichten diese Streckungsfaktoren jedoch nicht aus. Deshalb wurden die Pfaderweiterungen (b) der Figur 8 angewendet. Diese dürfen jedoch nur verwendet werden, wenn die Distanzmatrix eine gewisse Güte aufweist, wenn also für die Merkmale,

deren Gewichtung und für das Distanzmass eine gute Wahl getroffen wurde. Sonst kann der optimale Pfad unerwünschte Wege durch die Distanzmatrix einschlagen, die eine falsche zeitliche Anpassung der Sprachsignale zur Folge haben.

Die Pfaderweiterungen (b) aus Figur 8 erlauben Streckungsfaktoren im Bereich [1/3, 3]. Diese Streckungsfaktoren genügen für gesprochene Passagen, nicht aber für Sprechpausen. Für die korrekte Anpassung der Sprechpausen sind viel extremere Streckungsfaktoren nötig. Dies stellt mit den global gewählten Pfaderweiterungen aber ein grosses Problem dar. Denn solch liberale Pfadbedingungen, wie sie für die Anpassung von Sprechpausen benötigt werden, führen während den gesprochenen Passagen zu komplett falschen Pfaden. Die Lösung dieses Problems wird in folgendem Kapitel beschrieben.

4.4 Adaptive Dynamic Time Warping

Offensichtlich müssen Sprechpausen anders behandelt werden als gesprochene Abschnitte des Sprachsignals. Die Idee war, die eigentlich global definierten Kontinuitätsbedingungen des Dynamic Time Warpings lokal abzuändern. Die Kontinuitätsbedingungen wurden nicht mehr global gewählt, sondern sind signalabhängig. Dieses Verfahren wird hier als *Adaptive Dynamic Time Warping* bezeichnet.

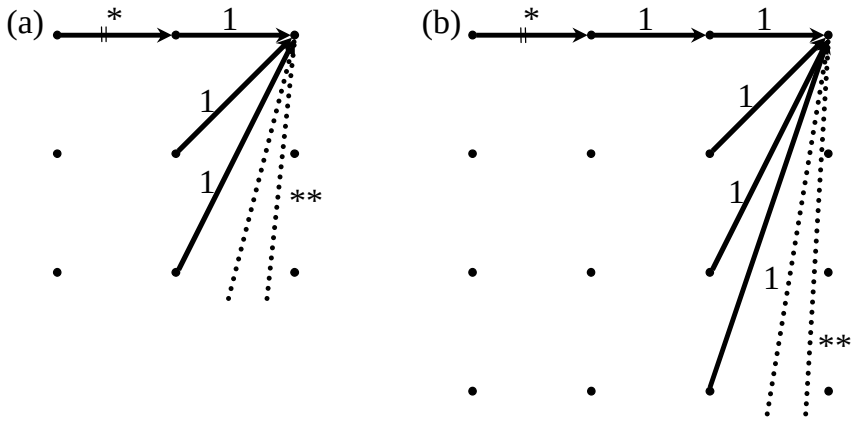
In Sprechpausen werden extremere Steigungen des Pfades verlangt. Während einer Sprechpause im synthetischen Signal (Merkmalssequenz X) ist eine potentielle Steigung von 0 nötig, während einer Sprechpause im natürlichen Sprachsignal (Merkmalssequenz Y) eine potentielle Steigung von ∞. Mit diesen zusätzlichen Bedingungen können Sprechpausen in beiden Signalen beliebig gestreckt oder gestaucht werden.

Die in dieser Arbeit eingesetzten Pfaderweiterungen sind in Figur 10 zu sehen. Es sind Modifikationen der Pfaderweiterungen aus dem vorangehenden Abschnitt. Die mit * und ** gekennzeichneten Pfade sind signalabhängig und unter einer Threshold-Bedingungen zugelassen. So ergeben sich in Abhängigkeit der Signalintensitäten insgesamt vier verschiedene Typen von Pfaderweiterungen. Es sind dies:

1. Pfaderweiterungen ohne Option * und **
2. Pfaderweiterungen mit Option *
3. Pfaderweiterungen mit Option **
4. Pfaderweiterungen mit Option * und **

Aus algorithmischen Gründen wurde die ideale Steigung von ∞ als eine maximal erlaubte Steigung von 10 verwirklicht.

Sprechpausen äussern sich durch eine geringe Intensität im Sprachsignal. Als Entscheidungskriterium wird also die Singalintensität jedes Frames herangezogen. Wenn nun die Intensität eines Frames unter dem Threshold TH_{int} liegt, ist an dieser Stelle in der Distanzmatrix eine extremere Steigung erlaubt. Genauer ausgedrückt, wenn das synthetische Signal unter dem Threshold liegt, ist der mit * gekennzeichnete Pfad erlaubt. Wenn das natürliche Sprachsignal unter dem Threshold liegt, sind die mit ** gekennzeichneten Pfade erlaubt. Bei Sprechpausen



Figur 10: *Zwei Typen von adaptiven, asymmetrischen Pfaderweiterungen (Adaptive Dynamic Time Warping)*

im synthetischen Signal ergeben sich dadurch in der Distanzmatrix vertikale Streifen, in denen zusätzlich die Steigung 0 über mehr als 2 Frames erlaubt ist. Bei Sprechpausen im natürlichen Sprachsignal ergeben sich in der Distanzmatrix horizontale Streifen, in denen eine zusätzliche maximale Steigung von 10 zugelassen wird.

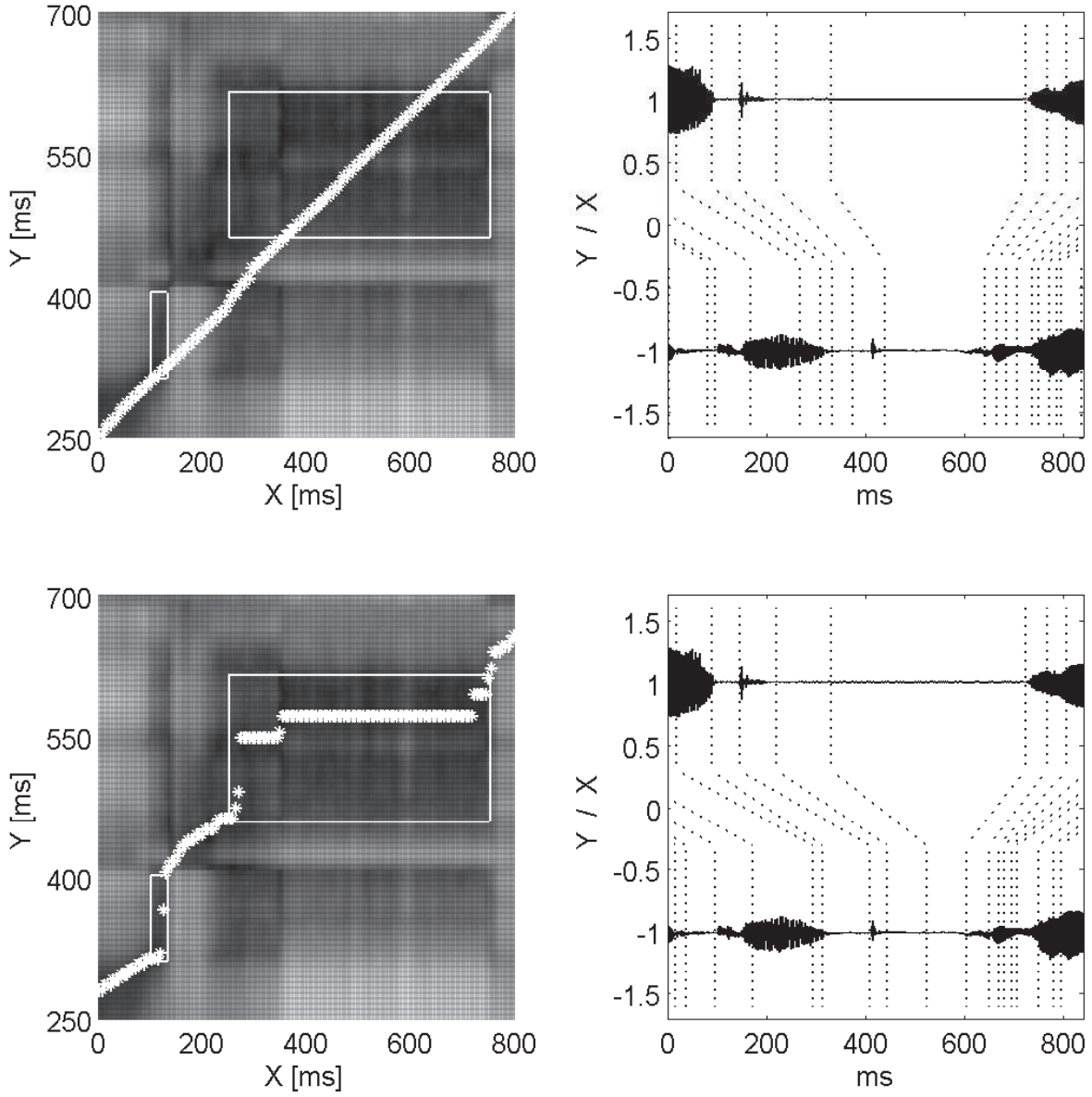
In Figur 11 ist der Vergleich zwischen Adaptive Dynamic Time Warping und Dynamic Time Warping anhand eines Beispiels illustriert. Im synthetischen Sprachsignal (oberes Signal) ist eine lange Sprechpause vorhanden, die im natürlichen Sprachsignal (unteres Signal) deutlich kürzer ist. In den Distanzmatrizen sind zwei Sprechpausen als weisse Boxen eingezeichnet. Der optimale Pfad von Dynamic Time Warping führt nicht durch die linken unteren und die rechten oberen Ecken dieser Boxen. Deshalb werden diese Pausen falsch angepasst, was einen Einfluss auf die Dauern der benachbarten Halbdiphon-Elemente hat. Im Gegensatz dazu verläuft der optimale Pfad von Adaptive Dynamic Time Warping dank den adaptiven Pfaderweiterungen quer durch diese Boxen. Die Ausschnitte werden dadurch richtig aneinander angepasst und die lange Sprechpause wird stark gekürzt.

4.5 Verifikation der zeitlichen Anpassung

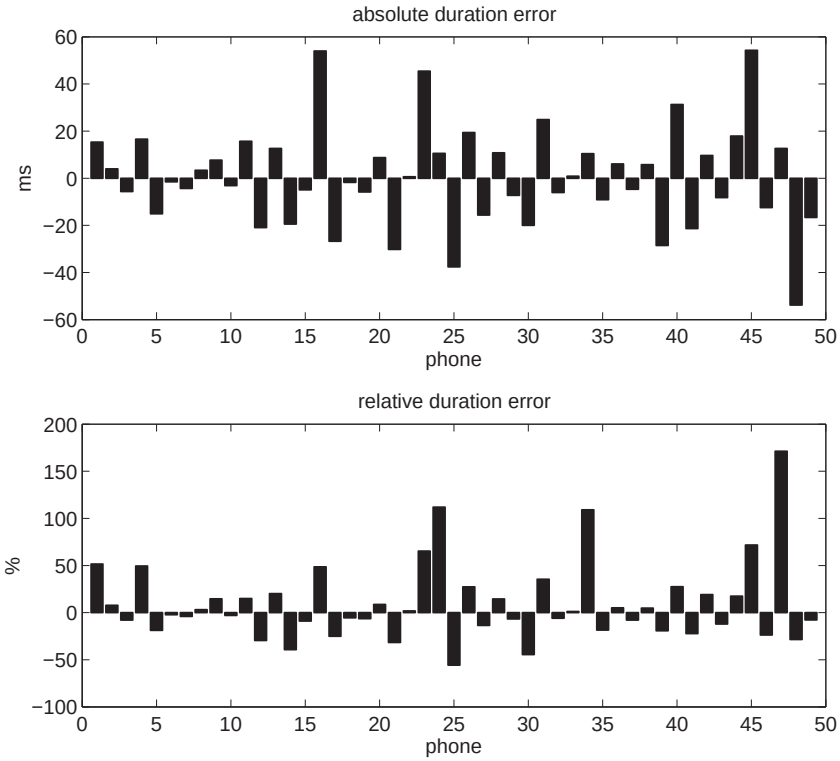
Bei der Anwendung von Dynamic Time Warping hat man mit der Wahl der Kontinuitätsbedingungen und des Distanzmasses eine Fülle von Parametern und Modifikationsmöglichkeiten. Dazu kommen die verwendeten Merkmale und deren Gewichtung. Umso mehr ist man auf eine robuste Verifikation des Algorithmus angewiesen. Ziel war es, die Güte und Korrektheit des angewendeten Dynamic Time Warpings, also der zeitlichen Anpassung der beiden Signale, zu untersuchen.

Es wurde ein manuelles Labeling der natürlichen Sprachsignale durchgeführt. Dabei wurden die Grenzen aller Halbdiphon-Elemente im Audiosignal von Hand ermittelt und markiert. Von einem Sprecher wurden zwei Varianten des ersten Satzes (siehe Kapitel 6.1) gelabelt. Von einem anderen Sprecher wurden die beiden Sätze 1 und 2 gelabelt. Insgesamt standen so die Labels von vier Aufnahmen zur Verfügung.

Der Labeling-Prozess ist sehr zeitaufwändig. Jeder verwendete Satz (siehe Kapitel 2.2.1)



Figur 11: *Vergleich von Dynamic Time Warping und Adaptive Dynamic Time Warping. Oben: Distanzmatrix mit optimalem Pfad und Halbdiphongrenzenübertragung aus Dynamic Time Warping. Unten: Distanzmatrix mit optimalem Pfad und Halbdiphongrenzenübertragung aus Adaptive Dynamic Time Warping*



Figur 12: *Beispiel der absoluten und relativen Lautdauer-Fehler eines aufgenommenen Satzes*

besteht aus 80 bis 120 Halddiphon-Elementen und benötigt gleich viele Labels. Um die Lage der Lautübergänge und Lautgrenzen abschätzen zu können, wurde das synthetische Sprachsignal mit den Halbdiphon-Grenzen aus der `synthele`-Datei als Referenz herangezogen. Zur besseren Erkennung von Vokalübergängen wurde zudem das Spektrogramm des natürlichen Signals eingesetzt. Neben der optischen Information diente auch das Anhören der Signalabschnitte zur Lokalisation der Halbdiphon-Grenzen. In einer zweiten Phase wurden die gelabelten Halbdiphon-Elemente nochmals akustisch beurteilt.

Problematisch war das Labeling vor allem dann, wenn im natürlichen Sprachsignal gewisse Laute gänzlich ausgelassen worden waren. In diesem Falle wurden die Labels der dazugehörigen Halbdiphone sehr kurz hintereinander gesetzt. Beispiele von ausgelassenen Lauten sind in Kapitel 6.5.4 angegeben.

Eine Schwierigkeit war zudem, in stimmhaften Lauten den Lautzentroiden zu finden. Vor allem bei langen, stimmhaften Lauten konnten die Labelgrenzen nur sehr ungenau gesetzt werden. Aber auch Lautübergänge wie z. B. der Übergang von einem stimmhaften Laut in einen anderen liessen sich oft nur schwer lokalisieren. Teilweise konnten die Labelgrenzen lediglich mit einer Genauigkeit von schätzungsweise $\pm 10ms$ gesetzt werden.

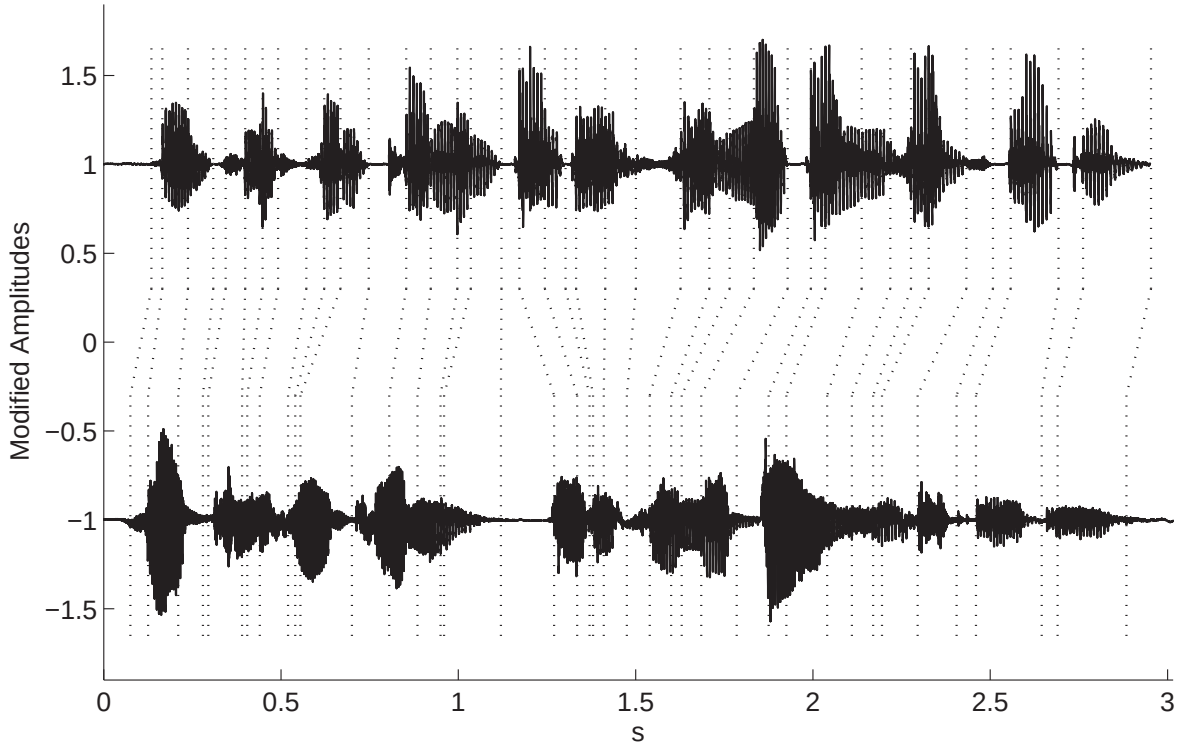
Durch das Labeling der Halbdiphon-Grenzen stand folglich auch die Information über die Halbdiphon-Dauer im natürlichen Signal als Referenz zur Verfügung. Diese durch das Labeling erhaltenen Referenz-Dauern wurden dann mit den Halbdiphon-Dauern verglichen, die durch die Dynamic Time Warping Algorithmen ermittelt worden waren. Es wurde die absolute mittlere Abweichung (ME) in ms sowie der absolute mittlere quadratische Fehler (MSE) dieser vier

Sätze berechnet. Die absoluten Fehler sagen jedoch nicht allzu viel über die wirkliche Qualität der Anpassung aus. Eine gleich grosse Abweichung erzeugt bei einem langen Halbdiphon und einem kurzen Halbdiphon denselben Fehler. Also wurden zusätzlich die relativen Fehler (ME und MSE) bezüglich der gelabelten Dauer ermittelt. In diesen relativen Fehlern fallen Abweichungen eines langen Halbdiphons weniger stark ins Gewicht als Abweichungen eines kurzen Halbdiphons. In Tabelle 6 sind die Fehler der vier erwähnten Aufnahmen angegeben. In Figur 12 sind die absoluten und relativen Fehler für jeden Halbdiphon eines aufgenommenen Satzes aufgezeigt. Bei den relativen Fehlern dominieren vor allem die Halbdiphon-Elemente von ausgelassenen Lauten.

Fehler	**Aufnahme 1**	**Aufnahme 2**	**Aufnahme 3**	**Aufnahme 4**
Absoluter ME $[ms]$	15.9	21.6	16.8	19.6
Absoluter MSE $[ms^2]$	442.4	773.4	467.5	806.7
Relativer ME $[\%]$	27.6	33.3	34.2	36.4
Relativer MSE $[\%^2]$	1799.6	1885.9	3643.3	3194.2

Tabelle 1: *Mittlere Fehler der vier gelabelten Aufnahmen*

Insgesamt betrachtet ist es sehr schwierig, ein gut anwendbares numerisches Qualitätsmerkmal für die Anpassung der beiden Signale zu finden. Die Ergebnisse aus den numerischen Analysen bestätigten aber die Vermutungen aus den akustischen und optischen Analysen der Sprachsignale. Dieser Umstand und die Tatsache, dass das Ergebnis für den Anwender subjektiv als gut wahrgenommen werden muss, rechtfertigen die schliesslich weiter angewendeten Verifikationsmethoden: Einerseits wurden die resynthetisierten Signale angehört und beurteilt, andererseits wurden optische Ausgaben der Anpassung studiert. In Figur 13 ist eine Lautzuordnung zweier Sprachsignale des Satzes `Herzlich willkommen bei der Firma Baumhaus AG` zu sehen. Das obere Signal entspricht dem synthetischen, das untere dem natürlichen Sprachsignal. Die Lautgrenzen der beiden Signale sind eingezeichnet. Die Lautgrenzen des natürlichen Signals wurden mittels Adaptive Dynamic Time Warping ermittelt. Mit etwas Kenntnis über das Aussehen von Sprachsignalen lassen sich diese Grenzen optisch verifizieren. Es war wichtig, sich mit den verwendeten Sprachsignalen intensiv auseinanderzusetzen, um die für die Algorithmen heiklen Passagen erkennen und genau überprüfen zu können.



Figur 13: *Darstellung der ermittelten Lautzuordnung zweier Sprachsignale*

5 Prosodiedetektion

5.1 Einleitung

Um die Prosodie des natürlichen Sprachsignals auf das synthetische Sprachsignal abbilden zu können, muss vorhergehend die Prosodie des natürlichen Sprachsignals bestimmt werden. Dauer, Grundfrequenzverlauf und Intensitätsverlauf müssen für jeden Laut ermittelt werden. Sobald diese Werte zur Verfügung stehen, ist die Prosodie des natürlichen Sprachsignals gegeben. Mit diesen Werten wird mit dem Sprachsynthesesystem polySVOX, das in Kapitel 2.2 beschrieben ist, das neue Sprachsignal erzeugt. Dazu werden die gefundenen Dauer- und Grundfrequenzstützwerte in eine `sigele`-Datei geschrieben und polySVOX als Input übergeben. So kann ein Sprachsignal mit der gewünschten Prosodie erzeugt werden.

5.2 Dauer

Die Dauerwerte des natürlichen Sprachsignals können einfach aus der Information der Anpassung der Signale gewonnen werden. Neben der Information der zeitlichen Anpassung wird die Annotation des synthetischen Sprachsignals benötigt. Wie in Kapitel 2.2 beschrieben, eignet sich dazu nur das `synthele`-Format. In dieser Datei sind die Halbdiphon-Elemente des synthetischen Sprachsignals inklusive deren Dauern enthalten. Die Übertragung der Start- und Endzeitpunkte sämtlicher Halbdiphon-Elemente vom synthetischen Sprachsignal auf das natürliche Signal liefert dann die Halbdiphon-Dauerwerte des natürlichen Sprachsignals.

Nicht so einfach ist die Übertragung von Sprechpausen. Im folgenden Kapitel wird die spezielle Behandlung von Sprechpausen erläutert.

5.2.1 Sprechpausen

Bei der Übertragung der Dauer gilt es, den Sprechpausen spezielle Beachtung zu schenken. Die Existenz und Länge von Sprechpausen variiert je nach Prosodie und Sprecher sehr stark. Für die subjektive Wahrnehmung der Betonung und die Verständlichkeit eines Textes sind solche Sprechpausen sehr relevant. Ein TTS-System kann einen langen Satz ohne Pause vorlesen, während ein menschlicher Sprecher zwingend eine Sprechpause, also eine Atempause, machen muss.

Problematisch ist die Übertragung der Sprechpausendauer vor allem dann, wenn im natürlichen Sprachsignal eine Sprechpause gemacht wurde, im synthetischen Signal aber an dieser Stelle keine Pause existiert. Dieser Fall tritt sehr häufig ein. Dementsprechend gibt es dann weder in der `phones`- noch in der `synthele`-Datei Pausenelemente. Ebenso fehlen die Halbdiphon-Elemente, die den Übergang in die Pause und den Übergang von der Pause in den nächsten Laut bilden. Dies verunmöglicht die Übertragung der Sprechpause als solche. Deshalb ist vor der eigentlichen Prosodieübertragung eine Pausendetektion mit gefolgter Pauseneinfügung erforderlich.

Die Pausendetektion basiert auf einem zusätzlichen, vorgängig durchgeführten Adaptive Dynamic Time Warping. Dieses Verfahren ist in Kapitel 4.4 erläutert. Dabei wurden genau die

gleichen Merkmale, Parameter und Kontinuitätsbedingungen verwendet wie für die anschliessende eigentliche Analyse der Prosodie. Im optimalen Pfad wird nach Stellen mit Steigungen grösser als TH_{slope} gesucht. Für TH_{slope} wurde der Wert 5 gewählt. Aufgrund der adaptiven Kontinuitätsbedingungen ist an diesen Stellen die Intensität des natürlichen Sprachsignals kleiner als der Threshold TH_{int}. Anschliessend wird in der Umgebung dieses Frames im synthetischen Signal nach Sprechpausen gesucht. Als Bereich wurden 50ms vor sowie nach dem eigentlichen Frame verwendet. Nur wenn keines der Frames in dieser Umgebung die Intensität TH_{int} überschreitet, wird eine Pause eingefügt.

Die so detektierten Pausen müssen anschliessend in die `phones`-Datei eingefügt werden. So werden die Übergänge in die Pause und von der Pause in den nächsten Laut von polySVOX generiert. Als Dauer einer eingefügten Pause wurde in dieser Arbeit stets 200ms gewählt. Die genaue Länge ist kaum relevant, da die Dauer der Pausen bei der darauffolgenden Prosodieübertragung noch verändert wird. Die Pausen wurden in dieser Arbeit manuell in die `phones`-Datei eingefügt. Aus dieser `phones`-Datei wurde dann von polySVOX ein neues synthetisches Sprachsignal generiert, welches dann die vorerst fehlenden Pausen enthielt.

5.3 Grundfrequenz

Damit die Sprechmelodie des natürlichen Sprachsignals übernommen werden kann, muss der Grundfrequenzverlauf dieses Signals bekannt sein. Zur Detektion des Grundfrequenzverlaufes gibt es verschiedene Ansätze, die zum Beispiel auf der Autokorrelation, dem Spektrum oder dem Cepstrum des Signals basieren. Komplexere Algorithmen kombinieren die Informationen aus den obengenannten Ansätzen miteinander. In dieser Arbeit wurde eine vom Institut für Technische Informatik der ETH Zürich zur Verfügung gestellte Funktion verwendet. Diese Funktion sucht zuerst den Grundfrequenzverlauf im Cepstrogramm. Der so ermittelte Verlauf wird im Spektrogramm verifiziert und allenfalls angepasst. Die maximal zu erkennende Grundfrequenz wurde bei Männerstimmen auf 400Hz, bei Frauenstimmen auf 800Hz beschränkt.

Für jedes Halbdiphon-Element wurden 5 äquidistante Stützwerte ermittelt, wobei die erste Stützstelle dem Startpunkt des Halbdiphons und die fünfte Stützstelle dem Endpunkt des Halbdiphons entsprach. Dann wurden aber nicht direkt diese Werte in die neue `sigele`-Datei geschrieben, denn so wäre dieser resynthetisierte Satz je nach Sprecher in einer anderen Stimmlage und würde nicht zu den anderen synthetisierten Sätzen der Anwendung passen. Es ist also eine relative Grundfrequenzübertragung erwünscht. Die ermittelten Grundfrequenzwerte müssen so mit einem Faktor korrigiert werden, dass der Mittelwert der neuen Grundfrequenzwerte demjenigen des ursprünglich synthetisierten Signals entspricht. So bleibt die Stimmlage der eingesetzten synthetischen Stimme erhalten.

5.4 Intensität

Nebst den Lautdauern und dem Grundfrequenzverlauf eines Sprachsignals gehört auch der Intensitätsverlauf zur Prosodie. In den meisten Sprachsynthesesystemen, die auf dem Verkettungsansatz basieren, wird die Intensität jedoch nicht gesteuert. Weil bei diesem Ansatz vorgesprochene Elemente verwendet werden, klingen die synthetisierten Sätze bereits ohne Intensitäts-

steuerung relativ gut.

Versuche, anschliessend zur Dauer- und Grundfrequenzübertragung auch den Intensitätsverlauf des natürlichen Sprachsignals zu übernehmen, scheiterten in dieser Arbeit. Der von polySVOX generierte Intensitätsverlauf wurde belassen. Die aufgetretenen Probleme werden in Kapitel 6.5.2 geschildert.

6 Experimente

6.1 Verwendete Sätze

Für die vorliegende Arbeit wurden vier Beispielsätze gebildet, welche mögliche Begrüssungstexte eines automatischen Systems sein könnten. Bei der Auswahl wurde darauf Wert gelegt, dass die Sätze mit möglichst verschiedenen Sprechweisen reproduziert werden können und sich die Texte in Wortschatz und der mittleren gesprochenen Länge unterscheiden. Zudem wurden bewusst imaginäre Firmennamen erfunden, deren erwünschte Betonung polySVOX nicht wissen kann. In der vorliegenden Arbeit wurden für die Analysen folgende vier Sätze verwendet:

```
Herzlich willkommen bei der Firma Baumhaus AG.

Ich begrüsse Sie beim automatischen Auskunftssystem
Fahrtvoraus.

Willkommen bei Multikom, der Lösung für alles.

Guten Tag. Sie sind hier beim Ticket-Reservationssystem
vom Kino Rex.
```

6.2 Aufnahmebedingungen

Die in dieser Arbeit verwendeten Audio-Daten wurden bewusst nicht unter Studiobedingungen aufgezeichnet. Ziel war es, eine der späteren Anwendung ähnliche Umgebung zu schaffen. Die Daten stammen von zwei weiblichen und einem männlichen nicht professionellen Sprecher. Die Sprecher wurden nicht speziell instruiert und produzierten somit während den Aufnahmen natürliche Nebengeräusche wie Räuspern oder Einatmungsgeräusche. Die vier im vorangehenden Kapitel erwähnten Sätze wurden von allen Sprechern mindestens zweimal vorgesprochen, wobei sie möglichst verschiedene Sprechweisen einsetzen mussten. Insgesamt standen so 33 natürliche Sprachsignale für die Analysen zur Verfügung.

Die Sprachsignale wurden mit dem Headset-Mikrophon «SL-8731» der Marke «Speed Link» und der externen Soundkarte «transit» der Marke «MAudio» aufgezeichnet. Es handelt sich um eine einfache und günstig erhältliche Ausrüstung und keinesfalls um ein professionelles Aufnahmesystem. Das Signal wurde Mono mit einer Abtastfrequenz von 16kHz und einer Auflösung von 16 bit aufgezeichnet.

6.3 Synthetische Stimme

Die synthetischen Sprachsignale wurden mit dem Sprachsynthesesystem polySVOX synthetisiert. Die in Kapitel 6.1 angegebenen Sätze wurden sowohl mit einer männlichen als auch mit einer weiblichen synthetischen Stimme erzeugt. Für die Prosodieübertragung eines von einem männlichen Sprecher vorgesprochenen Satzes wurde die männliche synthetische Stimme verwendet. Bei den Sprecherinnen wurde die weibliche synthetische Stimme eingesetzt. Es wird

jedoch vermutet, dass die Prosodieübertragung von einer männlichen auf eine weibliche Stimme und umgekehrt mit dem vorgestellten Verfahren ebenfalls möglich ist. Ein Beispiel einer solchen Prosodieübertragung von einer natürlichen männlichen Stimme auf eine synthetische weibliche Stimme ist in [CD:T34] und [CD:T35] gegeben.

6.4 Resultate

Auf der beiliegenden Audio-CD sind einige Resultate der in diesem Bericht vorgestellten und implementierten Prosodieübertragung enthalten. Wie in Kapitel 5 beschrieben, wurde sowohl die Dauer der Halbdiphone als auch der Grundfrequenzverlauf des vorgesprochenen, natürlichen Sprachsignals auf das neu synthetisierte Sprachsignal übertragen. Diese Übertragungen konnten unabhängig voneinander ein- und ausgeschaltet werden. In Tabelle 2 sind die Trackreferenzen der Anpassungen von zwei verschiedenen Sprachsignalen aufgeführt. Für die Grundfrequenz wird die Abkürzung F0 verwendet.

	männl. Sprecher, Satz 1	**weibl. Sprecher, Satz 4**
synthetisches Signal	[CD:T01]	[CD:T06]
natürliches Signal	[CD:T02]	[CD:T07]
Übertragung der Dauer	[CD:T03]	[CD:T08]
Übertragung der F0	[CD:T04]	[CD:T09]
Übertragung der Dauer & F0	[CD:T05]	[CD:T10]

Tabelle 2: *Übertragung der Dauer, der Grundfrequenz (F0) und beider Komponenten*

Es fällt auf, dass die Übertragung von nur einer der beiden Komponenten noch zu keinem überzeugenden Ergebnis führt. Erst mit der Anpassung der Dauerwerte *und* des Grundfrequenzverlaufs wird die Prosodieübertragung als gut wahrgenommen. Man beachte auch die relative Übertragung der Grundfrequenz und vergleiche die Tonlage der resynthetisierten Signale der unteren drei Zeilen Tabelle 2 mit dem ursprünglich synthetisierten Signal aus der ersten Zeile. Vor allem bei der Männerstimme ist die Erhaltung der Tonlage gut hörbar.

Weitere Beispiele der Prosodieübertragung (Dauer und Grundfrequenz) sind in Tabelle 3 angegeben. Alle Beispiele enthalten den gleichen Satz. Beispiele 1–3 stammen von einem männlichen Sprecher und basieren auf dem gleichen synthetischen Sprachsignal, das modifiziert wird. Beispiele 4 und 5 stammen von einer Sprecherin. Ihnen liegt ein synthetisches Sprachsignal einer Frauenstimme zu Grunde. Diese Beispiele veranschaulichen die Variationsmöglichkeiten der Prosodie und deren Übertragung.

	Bsp. 1	**Bsp. 2**	**Bsp. 3**	**Bsp. 4**	**Bsp. 5**
synth. Signal	[CD:T11]	[CD:T11]	[CD:T11]	[CD:T18]	[CD:T18]
nat. Signal	[CD:T12]	[CD:T14]	[CD:T16]	[CD:T19]	[CD:T21]
resynth. Signal	[CD:T13]	[CD:T15]	[CD:T17]	[CD:T20]	[CD:T22]

Tabelle 3: *Fünf Beispiele der Prosodieübertragung (Dauer und Grundfrequenz) des Satzes 1*

Wie bereits in Kapitel 5.4 erläutert, wurde auch die Übertragung des Intensitätsverlaufs vom natürlichen Sprachsignal auf das synthetische Sprachsignal getestet. Die Beschreibung der aufgetretenen Probleme und Hinweise auf Hörbeispiele der beiliegenden CD sind in Kapitel 6.5.2 aufgeführt. Im folgenden Kapitel sind zudem andere Probleme und Herausforderungen der Prosodieübertragung erläutert.

6.5 Probleme

6.5.1 Utterance Detection

Im natürlichen Sprachsignal war die Erkennung der Äusserung eine Herausforderung. Als erster Versuch wurde im Sprachsignal vom Signalbeginn, respektive vom Signalende her kommend mit einem Intensitäts-Threshold nach dem Äusserungsbeginn, respektive dem Äusserungsende gesucht. Dieses Vorgehen erwies sich für die verwendeten Daten jedoch als ungenügend. Vor allem bei Störgeräuschen, die durch den Sprecher verursacht worden waren, traten Probleme auf. Störgeräusche wie zum Beispiel Ein- und Ausatmungsgeräusche, Zungenschnalzer und Räuspern treten oft vor oder unmittelbar nach einer Äusserung auf. Mit solchen Störgeräuschen muss in der Anwendung auf jeden Fall gerechnet werden.

Aus diesem Grund musste zur Detektion des Start- und Endzeitpunktes der Äusserung ein komplexeres Verfahren angewendet werden. Dieses ist in Kapitel 4.2.1 und 4.2.2 beschrieben. Mit diesem Verfahren konnten Störgeräusche vor und nach der effektiven Äusserung weggeschnitten und das Problem somit gelöst werden.

6.5.2 Intensitätsübertragung

Der Intensitätsverlauf des natürlichen Sprachsignals wurde abschnittweise linearisiert und dann auf das resynthetisierte Sprachsignal übertragen. Das so erhaltene Signal klingt wenig überzeugend. Einerseits gibt es Probleme bei stark variierenden Stellen, wie zum Beispiel bei Glottalverschlüssen. Andererseits wird die Aussage im intensitätsangepassten Signal vor allem gegen Ende der Phrasen undeutlich und wirkt somit unprofessionell. Wieso die Aussage im natürlichen Sprachsignal mit demselben Intensitätsverlauf nicht als undeutlich wahrgenommen wird, kann nicht klar beantwortet werden. Der Grund dafür könnte unter anderem damit zusammenhängen, dass die vom Menschen wahrgenommene Lautstärke stark frequenzabhängig ist.

In Tabelle 4 sind die Trackhinweise für ein Beispiel mit der obengenannten Intensitätsübertragung angegeben. In [CD:T25] können die beschriebenen negativen Effekte angehört werden.

natürliches Signal	[CD:T23]
Übertragung von Dauer & F0	[CD:T24]
Übertragung von Dauer, F0 & Intensität	[CD:T25]

Tabelle 4: *Beispiel der Intensitätübertragung*

Um eine qualitativ überzeugende Intensitätsübertragung realisieren zu können, wären mehr

Analysen und Ansätze nötig. Ob die Prosodieübertragung mit einer zusätzlichen Intensitätsübertragung subjektiv verbessert werden kann, ist unklar.

6.5.3 Sprechpausen

Unerwarteterweise ergaben sich im Zusammenhang mit Sprechpausen gleich zwei Probleme. Einerseits variieren die Dauern von Sprechpausen sehr stark. Sie verlangen deutlich extremere Streckungsfaktoren als gesprochene Abschnitte des Sprachsignals. Dieses Problem wurde mit einer Modifikation des Dynamic Time Warpings gelöst. Die eigentlich global definierten Pfaderweiterungen wurden signalabhängig, also adaptiv, gemacht. Daraus resultiert das in Kapitel 4.4 besprochene Adaptive Dynamic Time Warping. Andererseits gibt es vor allem in langen Sätzen Fälle, in denen im natürlichen Sprachsignal im Gegensatz zum synthetischen Sprachsignal eine Sprechpause gemacht wird. In diesen Fällen ist eine korrekte zeitliche Anpassung wegen den im synthetischen Sprachsignal fehlenden Pausenelementen sowie den an die Sprechpause grenzenden Halbdiphon-Elementen nicht möglich. Deshalb wird vor der eigentlichen Prosodieübertragung eine Pausendetektion mit gefolgter Pauseneinfügung durchgeführt. In Tabelle 5 sind die Anspielhinweise eines Beispielssatz einer solchen Pauseneinfügung gegeben.

synthetisches Signal	[CD:T26]
natürliches Signal	[CD:T27]
resynthetisiertes Signal (ohne Pause)	[CD:T28]
modifiziertes, synthetisches Signal	[CD:T29]
resynthetisiertes Signal (mit Pause)	[CD:T30]

Tabelle 5: *Beispiel der Sprechpauseneinfügung*

Ohne Pauseneinfügung wird im resynthetisierten Sprachsignal [CD:T28] das Halbdiphon-Element vor sowie nach der eigentlichen Sprechpause über die ganze Dauer der Sprechpause gestreckt. Erst mit der Pauseneinfügung kann an der gewünschten Stelle im resynthetisierten Sprachsignal [CD:T30] die gewünschte Sprechpause verwirklicht werden. Dieses Problem kann also als gelöst betrachtet werden.

6.5.4 Ausgelassene Laute

Sehr oft werden von nicht professionellen Sprechern Laute unbewusst ausgelassen. Drei Wort-Beispiele für das Auslassen von Lauten sind in Tabelle 6 angegeben.

Originalwort	**effektiv gesprochenes Wort**
automatischen	autmatschn
Reservation	Resvation
Willkommen	Wikommn

Tabelle 6: *Wort-Beispiele für das Auslassen von Lauten*

Ausgelassene Laute erschweren die zeitliche Anpassung des natürlichen und des synthetischen Sprachsignals. Die vorgestellten Algorithmen können bis zu einem gewissen Grad mit ausgelassenen Lauten umgehen. Im natürlichen Sprachsignal in Figur 13 wurde im Wort kommen der Laut [ə] ausgelassen. Wie beim Zeitpunkt $0.9s$ zu sehen ist, wurde diese Lautauslassung vom Adaptive Dynamic Time Warping Algorithmus richtig erkannt. Werden jedoch zu viele Laute hintereinander ausgelassen, wird die zeitliche Anpassung ungenau.

Neben der Erschwerten zeitlichen Anpassung gibt es noch ein grundlegenderes Problem. Von der synthetischen Stimme wird eine korrekte Aussprache verlangt. Im resynthetisierten Sprachsignal dürfen also keine Laute ausgelassen werden. Textpassagen, die von der synthetischen Stimme auf Grund von ausgelassenen Lauten sehr schnell ausgesprochen werden, wirken aufgrund der schnellen und doch perfekten Artikulation künstlich. Um dieses Problem zu lösen, wäre eine Definition von Mindestdauern für Laute erforderlich. Solche Dauergrenzen wurde im Rahmen dieser Arbeit nicht gesetzt.

6.5.5 Knarrer

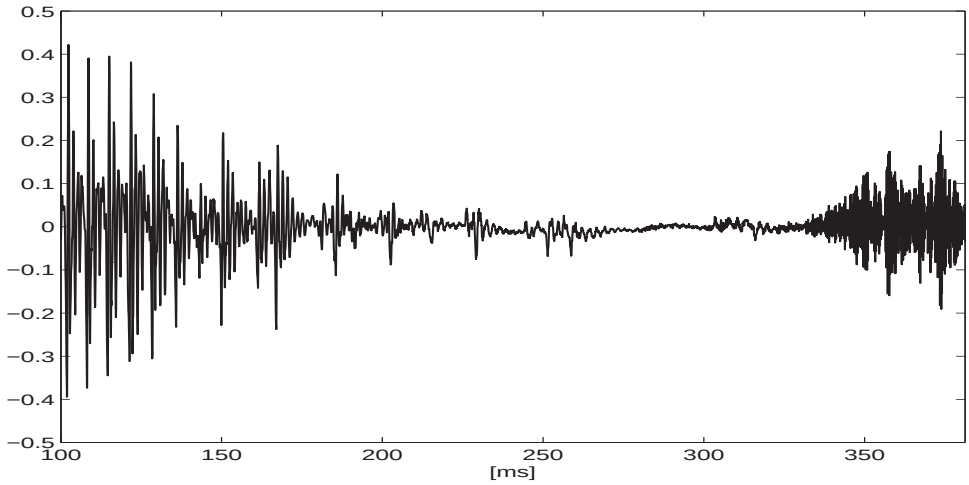
Bei der Männerstimme ergaben sich bei knarrend ausgesprochenen Passagen Probleme bei der Grundfrequenz-Detektion. Diese Knarrer, welche meist bei Wortanfängen oder Wortendungen auftreten können, verunmöglichten die Grundfrequenzdetektion. In Figur 14 ist ein solcher Knarrer im Sprachsignal eines männlichen Sprechers zu sehen. Der Knarrer beginnt ungefähr beim Zeitpunkt 150ms und endet bei 300ms. Man beachte vor allem die grosse Periodendauer im Vergleich zum vorangehenden Laut [u]. In Tabelle 7 sind die Anspielhinweise für die CD angegeben. Der erwähnte Knarrer ist im letzten Wort zwischen dem letzten und dem vorletzten Laut zu hören. Im resynthetisierten Signal wird an dieser Stelle die Grundfrequenz konstant gehalten, was nicht falsch klingt, aber nicht der vorgesprochenen Prosodie entspricht.

natürliches Signal	[CD:T31]
synthetisches Signal	[CD:T32]
resynthetisiertes Signal	[CD:T33]

Tabelle 7: *Beispiel eines Knarrers und dessen Auswirkung auf die Grundfrequenzübertragung*

6.5.6 Glottalverschlüsse

Es wird vermutet, dass Glottalverschlüsse bei der Prosodieübertragung ebenfalls Probleme verursachen könnten. Glottalverschlüsse können direkt nach einer Sprechpause auftreten. Wird im synthetischen Sprachsignal eine Pause eingefügt, muss überprüft werden, ob zusätzlich zur Pause ein Glottalverschlusselement in die phones-Datei eingefügt werden muss. Bei den untersuchten Daten war eine solche Einfügung aus akustischer Sicht nicht nötig. Dieses mögliche Problem wurde nicht weiter untersucht.



Figur 14: *Typischer Knarrer im Übergang vom Laut [u] zum Laut [s]*

Fazit

Das *Ziel* dieser Arbeit war es, Algorithmen für die Erweiterung eines bestehenden Sprachsynthesesystems zu entwickeln, mit der eine benutzerfreundlichen Eingabe der gewünschten Prosodie möglich wird. Dieses Ziel konnte mit dem verfolgten Ansatz erreicht werden. Mit dem vorgestellten Verfahren kann die gewünschte Prosodie dem Sprachsynthesesystem vorgesprochen werden, worauf dieses die natürliche Prosodie auf das synthetisiertes Sprachsignal überträgt.

Die *Prosodieübertragung* konnte mit der Übertragung der Halbdiphondauern und des Grundfrequenzverlaufs grösstenteils erreicht werden. Der Intensitätsverlauf des natürlichen Sprachsignals konnte jedoch nicht überzeugend auf das synthetische Sprachsignal übertragen werden.

Die *Hauptherausforderung* der Arbeit war die zeitliche Anpassung des natürlichen und des synthetischen Sprachsignals. Diese zeitliche Anpassung wurde mit einer Erweiterung des Dynamic Time Warping Verfahrens erreicht. Statt der global definierten Pfaderweiterungen wurden adaptive Pfaderweiterungen eingesetzt. Die Verifikation der zeitlichen Anpassung konnte nicht quantitativ erfasst werden. Die Resultate der Prosodieübertragung wurden durch das Anhören der Sprachsignale qualitativ verifiziert.

Das *Hauptproblem* der Prosodieübertragung sind im natürlichen Sprachsignal ausgelassene Laute. Diese erschweren einerseits die zeitliche Anpassung der Sprachsignale und andererseits verunmöglichen sie die genaue Prosodieübertragung auf ein korrekt ausgesprochenes, synthetisches Sprachsignal. Der Anwender der Sprachsynthesesystems sollte sich also bemühen, bei der Eingabe der gewünschten Prosodie möglichst keine Laute auszulassen und so den gewünschten Satz korrekt auszusprechen.

Ein *unerwartetes Problem* bei der Prosodieübertragung waren die Sprechpausen, die stark in ihrer Länge variieren. Dieses Problem wurde mit der Verwendung einer Pausendetektion sowie der Einführung adaptiver Pfaderweiterungen für das Dynamic Time Warping gelöst.

Das in der *Aufgabenstellung* erwähnte neuronale Netz wurde nicht eingesetzt. Aus zeitlichen Gründen musste zudem auf eine Beurteilung der Prosodieübertragung durch Testpersonen verzichtet werden.

Ausblick

Primär müssen mehr *Experimente* mit der vorgestellen Prosodieübertragung durchgeführt werden. Es sollen mehr Sätze, aber vor allem auch mehr Sprecher eingesetzt werden. Die Sprecher sollen die prosodieangepassten Sprachsignale subjektiv beurteilen und bewerten. Aus den Bewertungen können Schwachstellen der Prosodieübertragung ausgemacht werden.

Für *ausgelassene Laute* muss eine Mindestdauer definiert werden, damit die synthetischen Sätze korrekt ausgesprochen werden und nicht aufgrund zu schneller Artikulation künstlich wirken. Diese Mindestdauer muss wahrscheinlich lautabhängig gewählt werden.

Mit einem *besseren Distanzmass* zur Berechnung der Distanzmatrix können extremere Steigungen des optimalen Pfades zugelassen werden. So können ausgelassene Laute besser erkannt werden. Als Distanzmass bietet sich insbesondere die Verwendung eines neuronalen Netzes an. Dieses muss vorgängig mit geeigneten Daten trainiert werden.

Zusätzlich zur Übertragung der Halbdiphon-Dauern und des Grundfrequenzverlaufes wäre die *Übertragung des Intensitätsverlaufs* denkbar. Dazu müssen zuerst Analysen und Ansätze ausgearbeitet werden. Ob die Prosodieübertragung mit der zusätzlichen Übertragung des Intensitätsverlaufs subjektiv verbessert werden kann, ist jedoch unklar.

Literatur

[GM05] Pfister B. Gerber M. Quasi Text-Independent Speaker Verification with Neural Networks. Technical report, Speech Processing Group - Computer Engineering and Networks Laboratory - ETH Zurich, 2005.

[KE01] Pazzani M. Keogh E. Derivative Dynamic Time Warping. In *First SIAM International Conference on Data Mining*, 2001.

[PB05] Beutler R. Pfister B. *Skript zur Vorlesung Sprachverarbeitung I*. Institut für technische Informatik und Kommunikationsnetze, ETH Zürich, 2005.

[Rom06] H. Romsdorfer. *polySVOX TTS Synthesis - User Manual*, 1.0 edition, October 2006.

[Tra95] C. Traber. *SVOX: The Implementation of a Text-To-Speech System for German*. PhD thesis, Computer Enigneering and Networks Laboratory, ETH Zürich, March 1995.

Anhang A

Institut für
Technische Informatik und
Kommunikationsnetze

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

Wintersemester 2006/07
(SA-2007-08)

SEMESTERARBEIT

für

Herrn Reto Pieren

Betreuer: M. Gerber
Stellvertreter: Dr. B. Pfister und H. Romsdorfer

Ausgabe: 23. Oktober 2006
Abgabe: 2. Februar 2007

Sprachsynthese mit spezieller Prosodie

Einleitung

Die Prosodiesteuerung von Sprachsynthesesystemen wird in der Regel so konzipiert, dass die Systeme sich für das neutrale Vorlesen von Texten eignen. Die Sprachsynthese produziert also eine Sprache, wie wir sie etwa von einem Nachrichtensprecher gewohnt sind. In manchen Anwendungen von Sprachsynthese kommt es jedoch vor, dass für einzelne Sätze eine andere Sprechweise gewünscht wird. Dies kann z.B. bei einem automatischen Auskunftssystem der Fall sein, wo der Benutzer freundlich und einladend begrüsst werden soll. Selbstverständlich muss die automatisch erzeugte Sprechweise der verwendeten Sprachsynthese für die meisten Ausgaben des Auskunftssystems passend sein, sonst wird besser eine andere Sprachsynthese gewählt.

Für den Ingenieur, der ein solches Auskunftssystem realisieren will, ist es normalerweise nicht möglich, die Prosodiesteuerung für einzelne Sätze so zu verändern, dass die gewünschte Sprechweise resultiert. Das Problem ist in erster Linie, dass der Ingenieur gar nicht über das notwendige Wissen über die Prosodie der Sprache verfügt, um die Sprechweise gezielt modifizieren zu können.

Grundsätzlich wäre es zwar möglich, ein Sprachsynthesesystem mit gewissen Eingriffsmöglichkeiten zu versehen, sodass der Anwender (im vorliegenden Fall also der Ingenieur)

mit der Prosodiesteuerung etwas pröbeln könnte. Ob aber auf diesem Weg das gewünschte Resultat überhaupt erreichbar ist, ist doch sehr zu bezweifeln.

Eine sinnvollere Alternative dazu ist, die Sprachsynthese so zu erweitern, dass die Prosodie für einen einzelnen Satz, also beispielsweise für die Begrüssung, anhand eines Sprachsignals vorgegeben werden kann, das in der passenden Weise von einer beliebigen Person gesprochen worden ist.

Verfahren zur Prosodieübertragung

In dieser Arbeit geht es nun darum, ein Verfahren zu entwickeln, welches das TIK Sprachsynthesesystem SVOX so erweitert, dass dem System die Prosodie eines Satzes anhand eines natürlichen Sprachsignals mit demselben Wortlaut vorgegeben werden kann. Grundsätzlich sind verschiedene Verfahren denkbar, um für einen Text T ein synthetisches Sprachsignal S_{s2} zu erzeugen, welches die Prosodie eines natürlichen Sprachsignals S_n aufweist. Es soll hier jedoch eine Methode zur Prosodieübertragung angewandt werden, welche die folgenden Schritte umfasst:

a) **Synthese von S_{s1}:** Aus dem Text T wird mit dem Sprachsynthesesystem SVOX das Sprachsignal S_{s1} mit der normalen synthetischen Prosodie erzeugt. Nebst dem Sprachsignal S_{s1} liefert das Sprachsynthesesystem auch die Information[1], aus welchen Diphonen S_{s1} zusammengesetzt worden ist und wo die Diphon- und die Lautgrenzen in S_{s1} liegen.

b) **Ermitteln der Warping-Kurve:** Mittels DTW wird die optimale zeitliche Zuordnung zwischen den Signalen S_n und S_{s1} gesucht. Dazu müssen aus den Sprachsignalen geeignete Merkmalssequenzen extrahiert werden, und für den DTW-Algorithmus werden zweckmässige Pfaderweiterungen und ein Distanzmass (z.B. euklidische Distanz, siehe [1]) gebraucht.

c) **Segmentierung von S_n:** Anhand der Warping-Kurve werden die Diphon- und die Lautgrenzen in S_n ermittelt. Dadurch ergibt sich die Dauer jedes Halbdiphons.

d) **Grundfrequenzdetektion in S_n:** Mit einem geeigneten Verfahren wird pro Halbdiphon die mittlere Grundfrequenz bestimmt.

e) **Synthese von S_{s2}:** Schliesslich kann mit der Halbdiphonsequenz und den aus S_n ermittelten Dauer- und Grundfrequenzwerten eine neue Sigele-Datei (siehe [2]) erzeugt und mit SVOX das Sprachsignal S_{s2} generiert werden.

Das so erzeugte Sprachsignal S_{s2} weist somit die gleiche Sprechmelodie und denselben Sprechrhythmus auf wie das natürliche Sprachsignal S_n.

[1] Die Prosodiesteuerung des Sprachsynthesesystems SVOX liefert für ein zu synthetisierendes Sprachsignal eine Liste von Halbdiphonen je mit der Angabe von Dauer und Grundfrequenz. Diese Information wird optional auf eine sogenannte Sigele-Datei ausgegeben. Umgekehrt kann auch eine Sigele-Datei als Eingabe benutzt werden, zu der dann das entsprechende Sprachsignal erzeugt wird.

Aufgabenstellung

In dieser Semesterarbeit ist das oben skizzierte Verfahren zu Prosodieübertragung zu verwirklichen und zu testen. Es wird das folgende Vorgehen vorgeschlagen:

1. Machen Sie sich mit dem Sprachsynthesesystem SVOX und den in dieser Arbeit zu verwendenden Ein- und Ausgabedateien vertraut (siehe Dokumentation [2]).
2. Stellen Sie das detaillierte Konzept für die Prosodieübertragung auf. Achten Sie insbesondere darauf, dass das Verfahren zweckmässig optimiert und getestet werden kann. Dabei sind einerseits die Korrektheit und Güte der Teilschritte (z.B. Präzision der Segmentierung von S_n oder der Grundfrequenzdetektion) zu evaluieren, anderseits interessiert auch wie gut das Verfahren die Sprechweise des Signals S_{s2} an S_n anzupassen vermag.
3. Besprechen Sie das Konzept mit dem Betreuer, nehmen Sie allenfalls nötige Anpassungen vor und setzen Sie es in ein Matlab-Programm um.
4. Optimieren Sie alle Parameter des Verfahrens und führen Sie geeignete Tests durch, mit denen die Qualität der einzelnen Verfahrensteile beurteilt werden kann. Es interessiert insbesondere, wie gut die Segmentierung und die Grundfrequenzbestimmung von S_n sind.
5. Fakultativ: Falls die Segmentierung S_n nicht genau genug ist, bietet sich die Möglichkeit an, für den Schritt b) anstelle der euklidischen Distanz eine optimierte Distanz einzusetzen, die mit einem entsprechend trainierten neuronalen Netz (NN) vom Typ Multi-Layer-Perzeptron realisiert werden kann. Überlegen Sie sich, wie gross das NN sein soll und wie Sie es trainieren können (siehe beispielsweise [3], http://www.tik.ee.ethz.ch/~spr/publ_spg.html). Besprechen Sie Ihre Überlegungen mit dem Betreuer. Implementieren Sie das NN-basierte Distanzmass und testen Sie ob sich die Präzision der Segmentierung erhöht hat.
6. Stellen Sie ein Set von Testsätzen zusammen und führen Sie einen kleinen subjektiven Test mit einigen Testpersonen durch. Überlegen Sie sich, was die Testpersonen zu beurteilen haben und in welcher Form diese ihr Urteil mitteilen sollen.

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

Literaturverzeichnis

[1] B. Pfister und R. Beutler. *Sprachverarbeitung I.* Vorlesungsskript für das Wintersemester 2005/2006, Departement ITET, ETH Zürich, 2005.

[2] H. Romsdorfer. The polySVOX TTS Synthesis System. User Manual Version 1.0. Institut TIK, ETH Zürich, October 2006.

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

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

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

Zürich, den 23. Oktober 2006

Anhang B

phones-Datei

Die `phones`-Datei des Textes `Guten Tag` sieht folgendermassen aus:

```
\G\g   28.76  0:251 25:251 50:251 75:251 100:252
\G\u:  126.98 0:252 25:259 50:275 75:288 100:303
>      50.89  0:303 25:307 50:311 75:308 100:304
\G\t   16.91  0:304 25:303 50:301 75:300 100:298
\G\@   32.90  0:298 25:294 50:291 75:288 100:285
\G\n   61.16  0:285 25:282 50:276 75:268 100:261
>      63.58  0:261 25:253 50:250 75:250 100:249
\G\t   24.05  0:249 25:248 50:248 75:247 100:246
\G\a:  185.01 0:246 25:236 50:226 75:216 100:207
>      79.87  0:207 25:206 50:205 75:205 100:205
\G\k   47.87  0:205 25:205 50:205 75:205 100:205
```

sigele-Datei

Nachfolgend ist ein Ausschnitt aus einer `sigele`-Datei für den Text `Guten Tag` aufgeführt. In der ersten Spalte ist die Bezeichnung des Halbdiphon-Elements angegeben. Weitere Spalten enthalten die Dauer in Millisekunden und 5 Stützwerte des Grundfrequenzverlaufs für jedes Halbdiphon-Element.

```
\G\/g.b   female1 17.56  0:251 25:251 50:251 75:251 100:251
\G\gu:.a  female1 13.00  0:251 25:251 50:251 75:251 100:252
\G\gu:.b  female1 68.75  0:252 25:256 50:260 75:268 100:276
\G\u:t.a  female1 58.93  0:276 25:283 50:289 75:296 100:303
\G\u:t.b  female1 29.37  0:303 25:305 50:308 75:310 100:309
\G\t@.a   female1 34.93  0:309 25:307 50:305 75:301 100:298
\G\t@.b   female1 17.00  0:298 25:296 50:294 75:292 100:290
\G\@n.a   female1 13.87  0:290 25:289 50:287 75:286 100:285
\G\@n.b   female1 28.31  0:285 25:283 50:282 75:280 100:277
\G\nt.a   female1 32.00  0:277 25:274 50:269 75:265 100:261
\G\nt.b   female1 7.68   0:261 25:260 50:259 75:258 100:257
\G\ta:.a  female1 80.06  0:257 25:251 50:250 75:249 100:246
\G\ta:.b  female1 115.18 0:246 25:240 50:234 75:227 100:221
\G\a:k.a  female1 70.00  0:221 25:217 50:214 75:210 100:207
\G\a:k.b  female1 55.81  0:207 25:206 50:206 75:205 100:205
\G\k/.a   female1 75.00  0:205 25:205 50:205 75:205 100:205
```