Detektion und Elimination von Störgeräuschen bei Frikativlauten

Simon Simonet

Semesterarbeit SA-2008-30

Herbstsemester 2008

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: T. Ewender und S. Hoffmann

Verantwortlicher: Prof. Dr. L. Thiele

Inhaltsverzeichnis

Zusammenfassung

Störgeräusche in Sprachaufnahmen, welche als Grundlage für Sprachsynthese-Systeme dienen sollen, sind höchst unerwünscht. Da aber auch Störgeräusche beim Sprechen selber entstehen können, sind sie nie ganz vermeidbar.

In dieser Arbeit werden typische Störgeräusche wie sie in Frikativen vorkommen analysiert. Es wird dann ein Verfahren vorgestellt mit dem die Störgeräusche anhand von Merkmalen aus dem Spektrum und der Autokorrelation detektiert werden können, und die Detektionsmethode wird auf ihre Klassifizierungsleistung getestet. Abschliessend wird ein mögliches Verfahren zur Elimination der detektierten Störgeräusche vorgestellt.

1 Einleitung

Selbst bei Studioaufnahmen von professionellen Sprechern kommt es vor, dass Störgeräusche in den Aufnahmen vorhanden sind. Mit Störgeräuschen sind in diesem Fall nicht solche gemeint, die technischer Natur sind, wie Rauschen, Brummen oder Klicken. Vielmehr entstehen die Störgeräusche ungewollt beim Sprechen selber.

Im Alltag fallen uns sprechbedingte Störgeräusche bei einmaligem Hören kaum auf. Sollen solche Aufnahmen allerdings als Grundlage für die Sprachsynthese dienen, stellen sie ein erhebliches Problem dar.

Bei der Korpussynthese kann es in einem ungünstigen Fall dazu kommen, dass ein Störgeräusch nicht nur einmal zu hören ist, sondern unter Umständen gleich mehrmals in einem Satz. Spätestens bei der zweiten oder dritten Wiederholung wird man darauf aufmerksam. Sind ausserdem Veränderungen der prosodischen Grössen am Signal notwendig, wie bei der Diphonsynthese, werden die störenden Geräusche allenfalls noch verstärkt. Ein Detektionsverfahren, welches verhindert, dass solche Störgeräusche in den Signalkorpus eines Sprachsynthesesystems gelangen, wäre folglich sehr nützlich.

In dieser Arbeit sollen Störgeräusche analysiert werden, die in Frikativlauten vorkommen, insbesondere in der Untergruppe der Zischlaute ([f],[v],[s],[z],[ʃ],[ʒ]). Für die Störgeräusche sollen dann geeignete Merkmale gefunden werden, die sich für die Detektion eigenen. Basierend darauf soll ein Detektionsverfahren implementiert und auf seine Klassifizierungsleistung hin untersucht werden. Zuletzt soll ein Weg gefunden werden, die detektierten Störgeräusche aus dem Signal zu entfernen.

Die exakte Aufgabenstellung vom Institut ist im Anhang B zu finden.

2 Beschreibung des Signalkorpus

Als Grundlage für die Untersuchung dient eine bereits vorselektierte Sammlung von kurzen Aufnahmen einer französischen Sprecherin. Die Sammlung besteht aus 72 einzelnen Aufnahmen mit Längen zwischen 1.5 und 18 Sekunden. Die Gesamtdauer der Aufnahmen beträgt 386 Sekunden. Die Abtastfrequenz der Aufnahmen beträgt 44.1 kHz.

Von den Signalen lag eine automatische Lautsegmentierung vor. Während der Klassifikation der Frikative mussten bei einigen wenigen Segmenten die zeitlichen Grenzen noch von Hand korrigiert werden. Abgesehen davon wurde die Segmentierung unverändert übernommen.

In der Tabelle 1 ist die absolute Häufigkeit der im Signalkorpus vorkommenden Frikative zu finden.

Laut	[s]	[ʃ]	[z]	[ʒ]	[f]	[v]
Absolute Häufigkeit	173	22	52	43	35	72

Tabelle 1: *Vorkommen der Frikativlaute*

2.1 Klassifikation der Frikative

Für die Klassifikation der Signale wurden sämtliche Frikative mehrfach angehört, sowohl isoliert wie auch als Teil des umgebenden Sprachsignals. Die Klassifikation der Frikative erfolgte anhand der Art des Störgeräusches und der Stärke der Wahrnehmbarkeit.

Für die Stärke der Wahrnehmbarkeit wurde eine grobe Einteilung in folgende drei Klassen vorgenommen:

- unauffällig ($K0$): Frikative bei denen kein Störgeräusch wahrnehmbar ist
- wahrnehmbar ($K1$): Frikative mit wahrnehmbarem Geräusch
- störend ($K2$): Frikative mit störendem Geräusch

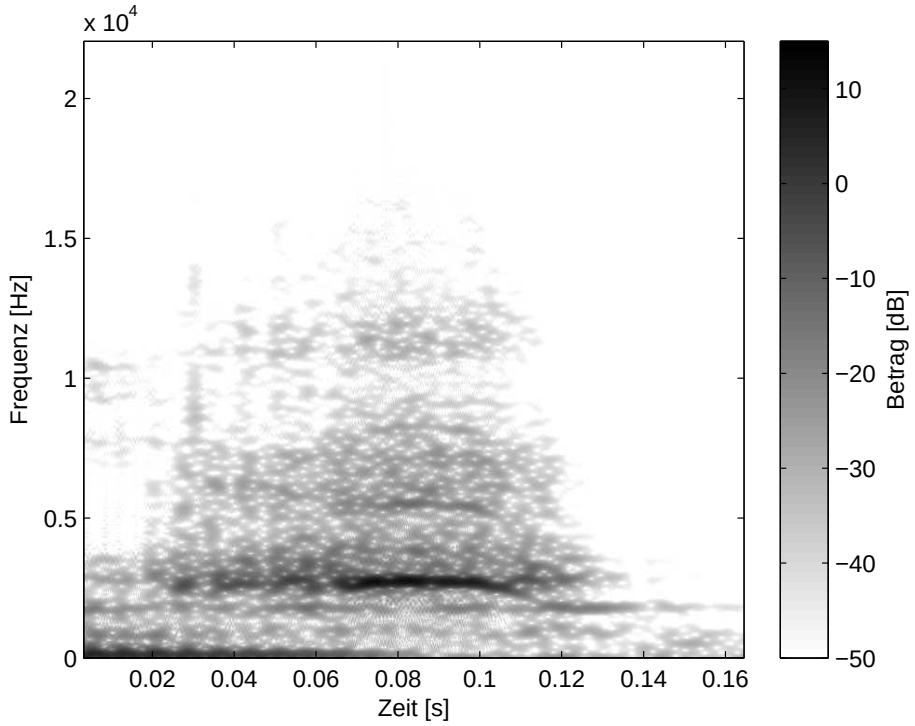
Den Klassen $K0$ und $K2$ wurden nur Frikative zugeordnet bei denen die Klassifikation eindeutig möglich war. Die Klasse $K1$ beinhaltet Frikative, bei denen eine gewisse Unsicherheit hinsichtlich ihrer Zuordnung bestand. Je nachdem wie kritisch man hinhört und abhängig davon, ob man nur den isolierten Laut oder den ganzen Satz hört, neigt man zu unterschiedlichen Beurteilungen über die Wahrnehmbarkeit eines eventuell vorhandenen Störgeräusches.

Desweiteren liessen sich im Wesentlichen vier unterschiedliche Geräuscharten in der Frikativsammlung ausmachen die im folgenden Abschnitt näher beschrieben werden sollen.

2.2 Störgeräuscharten

2.2.1 Pfeifen

Der akustisch am besten wahrzunehmende Geräuschtyp ist das Pfeifen. Das Pfeifgeräusch in den stark betroffenen Frikativen ist durchaus mit einem absichtlich erzeugten Pfeifton vergleichbar. Das Pfeifen ist besonders in den Sch-Lauten, sowohl in den stimmlosen als auch in den stimmhaften, und den f-Lauten anzutreffen. Im zeitlichen Verlauf des Spektrums, in Fi-



Figur 1: *Spektrogramm eines Frikativs mit typischem Pfeifgeräusch*

gur 1 in Form eines Schmallbandspektrogramms dargestellt, ist das Pfeifen gut an einem oder mehreren ausgeprägten Formanten erkennbar. Die dominanten Formanten liegen in einem Bereich zwischen 1600-5000 Hz. Formanten im Bereich von der Sprachgrundfrequenz aufwärts bis hin zu 1600 Hz sind auch bei den stimmhaften Lauten nur schwach ausgeprägt.

Die Frequenzen der einzelnen dominanten spektralen Überhöhungen im Frikativ ändern sich im Verlauf der Zeit nur wenig. Hingegen ist häufig zu beobachten, dass die Intensität mancher Formanten hin zur Lautmitte abnimmt und dann wieder ansteigt. Je kleiner die Schwankungen der Intensität sind, umso besser ist das Pfeifen wahrnehmbar.

Ebenfalls kann man feststellen, dass die Breite des Formanten einen Einfluss auf die Wahrnehmbarkeit des Störgeräusches hat. Schmalbandige Formanten gehen mit einem deutlich hörbaren Geräusch einher. Ausgeprägte lokale Minima in unmittelbarer Nähe zum lokalen Maximum verstärken ferner die Wahrnehmung des Störgeräusches. Gleiches gilt nicht nur für das Pfeifen, sondern auch für die anderen Geräuscharten.

Weiter kann man feststellen, dass der Einfluss dieser Faktoren sich mit zunehmender Fre-

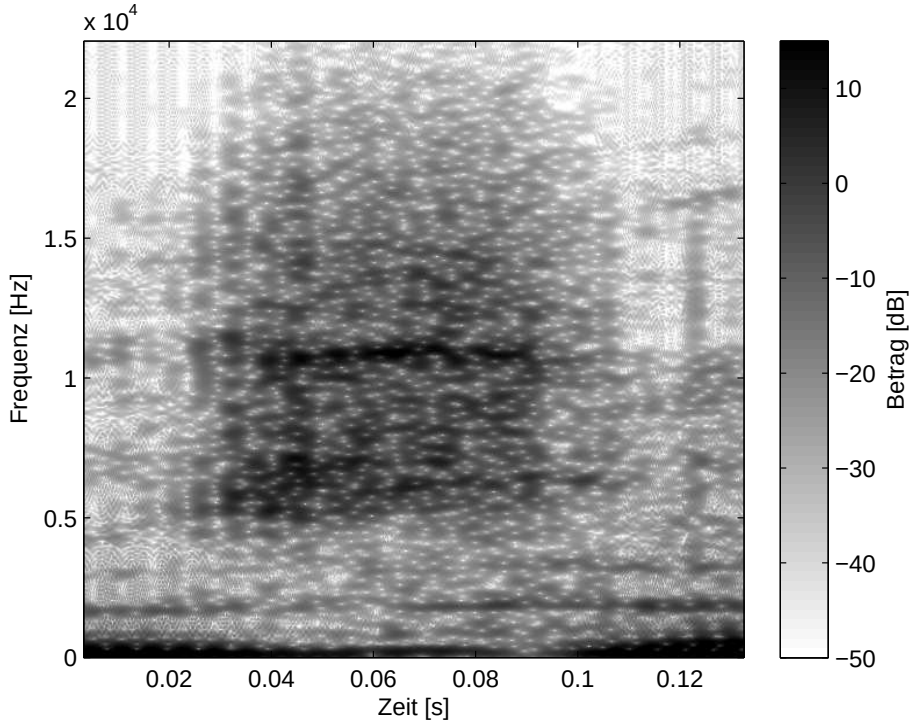
quenz abschwächt. So ist ein Pfeifgeräusch mit einem ausgeprägten Formanten bei 2 kHz im Allgemeinen besser wahrnehmbar als bei 5 kHz.

Die Länge bzw. Dauer, in der diese Struktur klar erkennbar ist, variiert von ungefähr zehn bis hin zu hundert Millisekunden und mehr. Akustisch lässt sich sowohl die zeitliche Position als auch die Dauer des Störgeräusches nur schwer schätzen.

Mit einem Equalizer wurde auf einfache Weise festgestellt, ob die Störgeräusche auch tatsächlich frequenzmässig dort zu finden sind, wo man sie nach der Betrachtung des Spektrogramms vermutet. Ein leichtes Verstärken des Frequenzbereichs hat dann auch entsprechenden Einfluss auf die Wahrnehmung des Störgeräusches.

2.2.2 Zischen

Die zischenden Geräusche ähneln klanglich den Geräuschen, die unter Druck stehende Gase beim Entweichen erzeugen, wie zum Beispiel der ausströmende Wasserdampf bei einem Schnellkochtopf. Die störenden Zischgeräusche kommen vorwiegend in den stimmlosen und den stimmhaften s-Lauten vor.

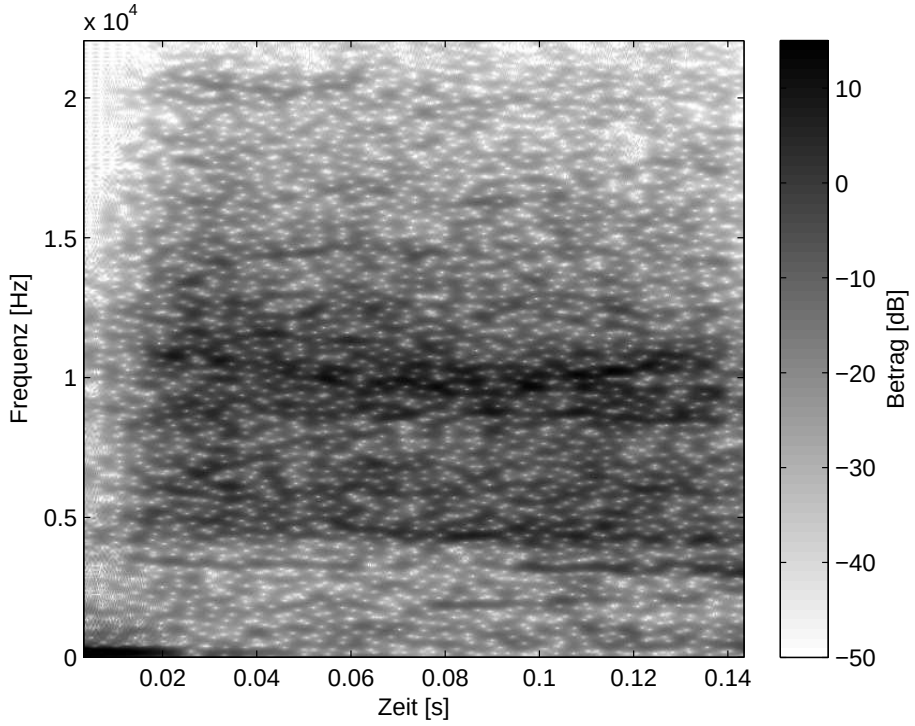


Figur 2: *Spektrogramm eines Frikativs mit typischem Zischgeräusch*

Im Schmalbandspektrogramm (Beispiel in Figur 2 dargestellt) der betroffenen Frikative erkennt man ein ähnliches Muster wie bei den Pfeifgeräuschen. Das dominierende spektrale Maximum ist allerdings in einem Bereich zwischen 9 bis 12 kHz zu finden. Ein zweiter, weniger ausgeprägter Formant befindet sich häufig in einem Bereich um die 5 kHz. Im Breitbandspektrogramm kann man auch beim Zischen beobachten, dass die Frequenzen der Formanten im Verlauf der Zeit etwa konstant sind.

2.2.3 Sirren

Das Sirren lässt sich am ehesten mit dem Geräusch eines fliegenden Bienenschwarms vergleichen. Allerdings ist der Klang deutlich heller, scharfer und hochfrequenter als beim Bienenschwarm. Störgeräusche mit dieser Klangcharakteristik findet man in den stimmlosen und stimmhaften s-Lauten.



Figur 3: *Spektrogramm eines Frikativs mit typischem Sirrgeräusch*

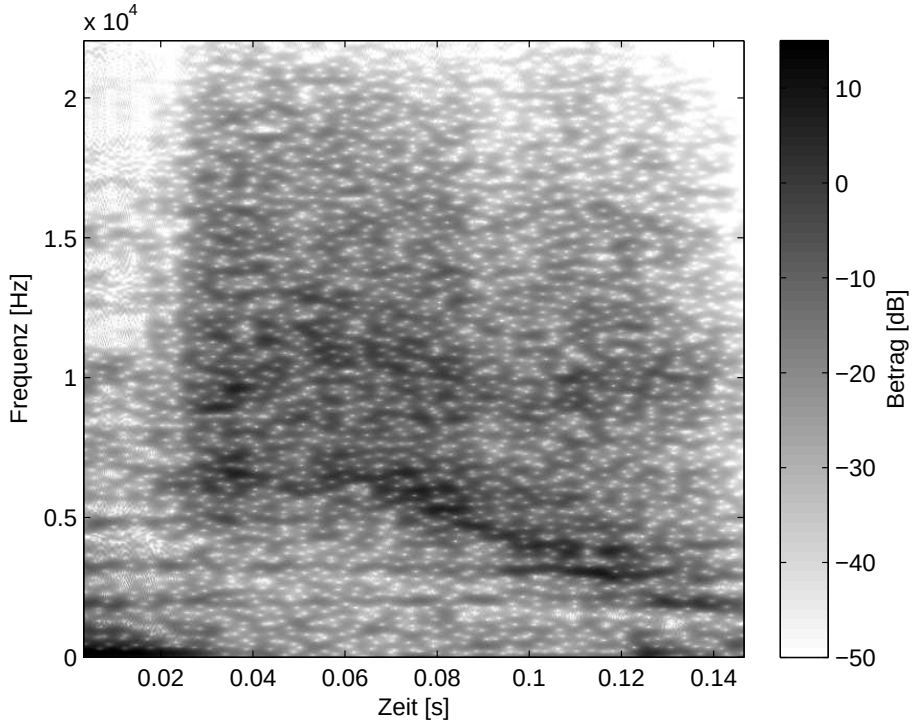
Die grobe Struktur des Spektrogramms von einem Frikativ mit Sirrgeräusch ähnelt derjenigen von einem Zischgeräusch. Es ist eine ausgeprägte Hochpasscharakteristik erkennbar mit maximalen Beträgen im Bereich um die 10 kHz. Im Gegensatz zu den Zischgeräuschen können aber meist mehrere, eng beieinander liegende Formanten ausgemacht werden, deren Amplituden im Verlauf der Zeit stärkeren Schwankungen unterworfen sind.

Akustisch werden diese nicht als einzelne Komponenten wahrgenommen, sondern machen sich nur als einen hochfrequenten Anteil bemerkbar. Man hört jedoch ein Schwanken der Intensität, was vermutlich der Grund für die Assoziation mit einem Sirren ist. Eine eindeutige Unterscheidung zwischen einem Sirren und Zischen gestaltet sich teilweise recht schwierig. Allgemeinen klingt das Zischen aber kompakter und präsenter als das Sirren.

2.2.4 Gleiten

Als vierten Geräuschtyp sei noch das Gleiten erwähnt. Klanglich ist das Geräusch mit einem Glissando auf einer Gitarre vergleichbar. Man hört demzufolge eine gleitende Veränderung der

Tonhöhe. Dieses Geräusch tritt am Übergang zwischen zwei Lauten auf. Im der Regel nimmt man es eher als unschönen Klang wahr und weniger als störendes Geräusch.



Figur 4: *Spektrogramm eines Frikativs mit typischem Gleitgeräusch*

Im Spektrogramm von Figur 4 sieht man das Gleiten des Betragsmaximums von ungefähr 7 auf 2.5 kHz. Ausserdem erkennt man 3 Formanten bei 2, 3.5 und 5 kHz die sich durch den Laut ziehen. Im Segment zwischen 3.48s und 3.5s nimmt die Intensität des mittleren Formanten stark zu. Damit einher geht auch ein hörbares Pfeifgeräusch. Diese Konstellation von Gleiten und einem Pfeifgeräusch ist auch bei weiteren Frikativen zu beobachten, insbesondere bei den Lautfolgen [si] und [sy]. Das Gleiten selber wird bei einem starken Pfeifgeräusch dann eher wenig wahrgenommen.

Die Ergebnisse der akustischen Klassifikation sind in der Tabelle 2 gelistet. Die Werte in den Klammern entsprechen der Anzahl der unterschiedlichen Störgeräuscharten (Pfeifen, Zischen, Sirren, Gleiten).

Klasse	$K0$	$K1$	$K2$
[s]	96	57 (7,20,23,7)	20 (8,6,3,3)
[z]	35	12 (2,5,5,0)	5 (3,2,0,0)
[ʃ]	0	1 (1,0,0,0)	21 (21,0,0,0)
[ʒ]	3	13 (13,0,0,0)	27 (27,0,0,0)
[f]	28	4 (4,0,0,0)	3 (3,0,0,0)
[v]	70	2 (2,0,0,0)	0 (0,0,0,0)
Gesamt	232	89	76

Tabelle 2: *Übersicht der akustischen Klassifikation der Frikative*

3 Wahl der Merkmale für die Detektion

Ausgehend von der Klassifikation der Frikative und den gewonnen Erkenntnissen über die spektralen Charakteristiken der Störgeräusche, wurden nun verschiedene Kurzzeit-Analysemethoden ausgewählt mit dem Ziel, geeignete Merkmale für die Detektion der Störgeräusche aus dem Sprachsignal zu extrahieren.

3.1 Autokorrelation

Die Autokorrelationsfunktion für ein zeitdiskretes und energiebegrenztes Signal ist definiert als

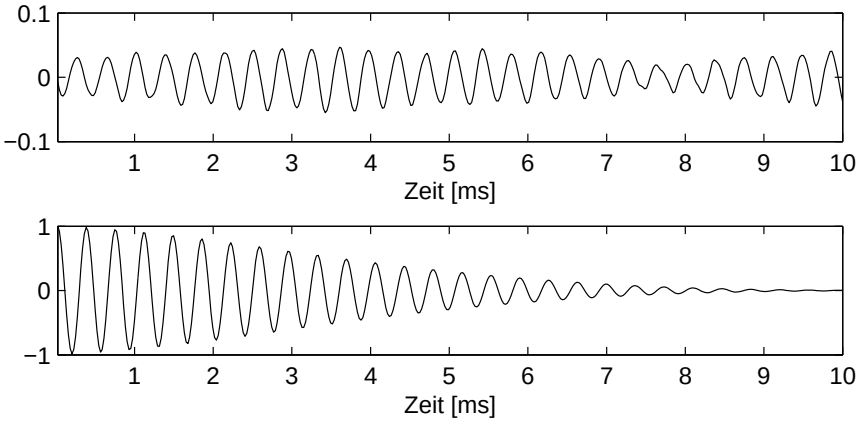
$$r(k) = \sum_{n=-\infty}^{\infty} x(n)x(n+k) \tag{1}$$

Der Wert der Autokorrelationsfunktion an der Stellen k beschreibt ein Mass für die Korrelation eines Signals mit dem zeitlich um k Abtastintervalle verschobenen, identischen Signal. Der Wert an der Stelle $k = 0$ entspricht der Energie des Signals.

Mit Hilfe der Kurzzeit-Autokorrelation, bei der die AKF nur über einen gefensterten Signalabschnitt berechnet wird, lassen sich so periodische Abschnitte im Signal finden wie sie bei Frikativen mit Störgeräusch beobachtet wurden.

Es ist ausserdem nützlich die Werte der AKF zu normieren, da die Absolutwerte in unserem Fall nicht erheblich sind. Hierfür werden die Werte der AKF durch den Wert an der Stelle $k = 0$ dividiert.

Aus den Positionen der relativen Maxima der Kurzzeit-Autokorrelation können die Periode bzw. die dazugehörige Frequenz als Merkmal extrahiert werden. Zu beachten ist hierbei, dass die Auflösung mit zunehmender Frequenz bzw. mit abnehmender Periodendauer geringer wird. Die Höhe der lokalen Maxima kann als Mass für die Stärke der Periodizität dienen.



Figur 5: *Signalabschnitt und normierte Kurzzeit-AKF des Abschnittes eines Frikativs mit Pfeifgeräusch*

In Figur 5 ist für ein Signalabschnitt aus einem Frikativ mit einem Pfeifgeräusch die normierte AKF dargestellt. Es handelt sich um den gleichen Frikativ wie in Abbildung 1. Aus dem

grössten lokalen Maximum, ohne das Maximum bei $k = 0$ zu berücksichtigen, lässt sich eine Periode von etwa 0.4 ms erkennen, was auf ein periodisches Signal mit der Frequenz 2.5 kHz schliessen lässt. Dies entspricht auch den im Spektrogramm gemachten Beobachtungen. Bei solch ausgesprochen stark periodischen Signalabschnitten ist ihre Periodizität auch ohne weiteres in der Wellenformdarstellung erkennbar.

3.2 Spektrale Merkmale

Die Charakteristiken der verschiedenen Störgeräusche lassen weiter vermuten, dass spektrale Merkmale bei der Detektion nützlich sein könnten. Da die Darstellung des Signals in einem Kurzzeit-Leistungsspektrum wesentlich mehr Informationen erhält als für die Merkmalextraktion nützlich ist, wurde hier ein Ansatz basierend auf der linearen Prädiktion verwendet. Die theoretische Abhandlung und Herleitung sind in [PK08] zu finden.

Die LPC-Analyse zerlegt das Sprachsignal in eine Sequenz von Filterkoeffizienten. Dabei nähert sich das LPC-Spektrum (der Betrag der Übertragungsfunktion des LPC-Filters) mit steigender Ordnung K des Prädiktors der Enveloppe des DFT-Spektrums an. Das LPC-Spektrum gibt somit den wesentlichen Verlauf des DFT-Spektrums wieder.

Der Vorteil vom LPC-Spektrum gegenüber einem DFT-Spektrum ist, dass die periodischen Anteile der glottalen Anregung nicht mehr in der Filterfunktion enthalten sind. Das Filter besitzt, abgesehen von einer p-Fachen Nullstelle bei $z = 0$, nur Pole und wird daher auch als Allpol-Filter bezeichnet. Deshalb werden Maxima im Spektrum allgemein besser approximiert als die Minima. Ebenfalls ist zu beachten, dass sehr schmale Maxima oft durch Pole mit einer zu hohen Güte approximiert werden. Für die Erkennung der im Spektrogramm festgestellten Charakteristiken der Störgeräusche eignet sich diese Näherung.

Aus den erhaltenen LPC-Spektren können folgende wesentliche Grössen der Maxima extrahiert werden:

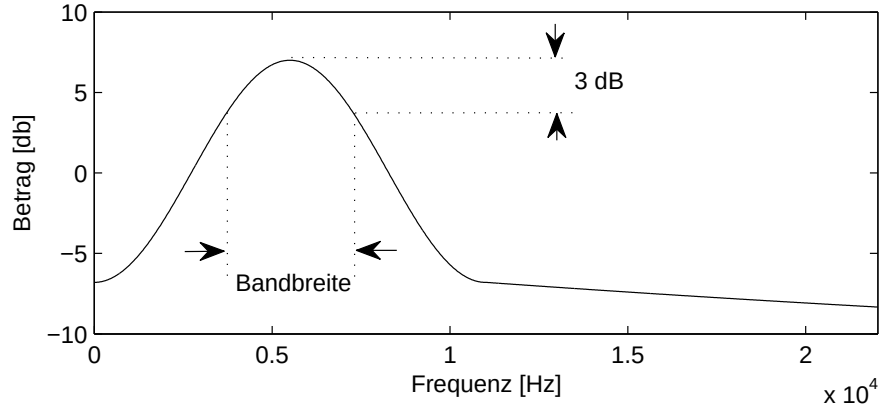
- Die Amplitude des Maximums
- Die Mittenfrequenz des Maximums
- Die Bandbreite des Maximums

Ausserdem werden auch die Amplituden und Frequenzen der Minima extrahiert und zuletzt der Betragsmittelwert des LPC-Spektrums.

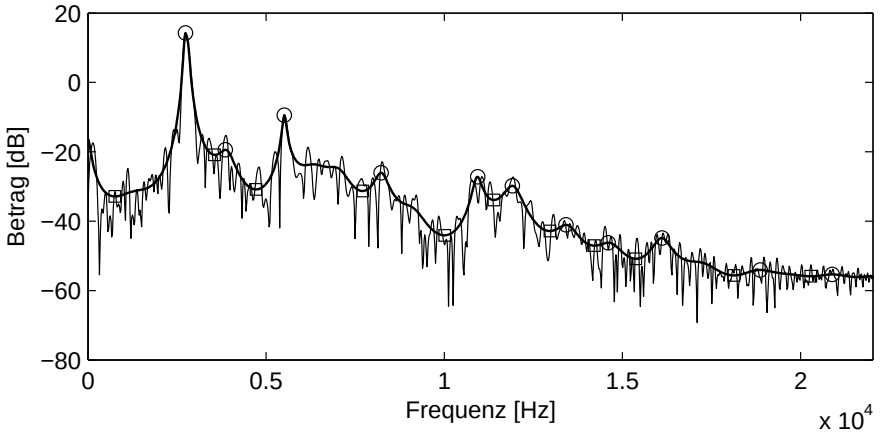
Die Bandbreite des Maximums ist hier definiert als Frequenzbereich zwischen dem Intensitätsmaximum und dem Intensitätsmaximum minus 3 dB, verdeutlicht in Figur 6.

In Abbildung 7 sind das DFT-Spektrum und das zugehörige LPC-Spektrum eines Analyseabschnittes vom Frikativ aus Abbildung 1 dargestellt. Die Kreise markieren die extrahierten Maxima, die Quadrate die lokalen Minima.

Ausgehend von den extrahierten Merkmalen aus einem LPC-Spektrum können nun durch betrachten mehrerer aufeinanderfolgenden Signalabschnitte und deren LPC-Spektren zusätzliche Informationen über den zeitlichen Verlauf der einzelnen Maxima ermittelt werden. Dabei sind insbesondere deren Frequenzverlauf und der Verlauf der Amplitude der Maxima von Interesse.



Figur 6: *Definition der Bandbreite eines Maximums im LPC-Spektrum*



Figur 7: *DFT-Spektrum und LPC-Spektrum eines Analyseabschnittes*

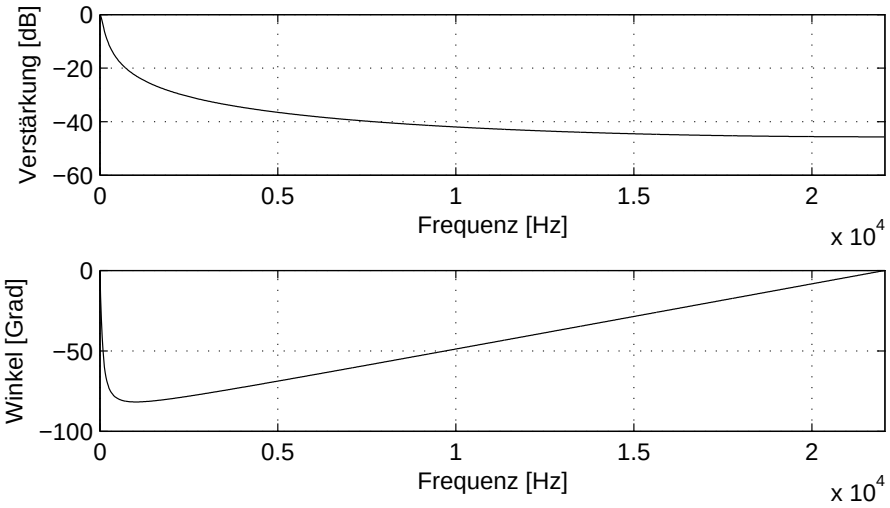
Als einfacher Ansatz wurde hierfür für ein Maximum aus dem Signalabschnitt n in den Signalabschnitten $n-k$ bis $n-1$ jeweils das frequenzmässig am nächsten gelegene spektrale Maximum gesucht. Aus den erhaltenen Werten der Maxima lassen sich so die Standardabweichung für Amplitude und Formantfrequenz als Mass für die zeitliche Varianz berechnen.

3.3 Signalenergie

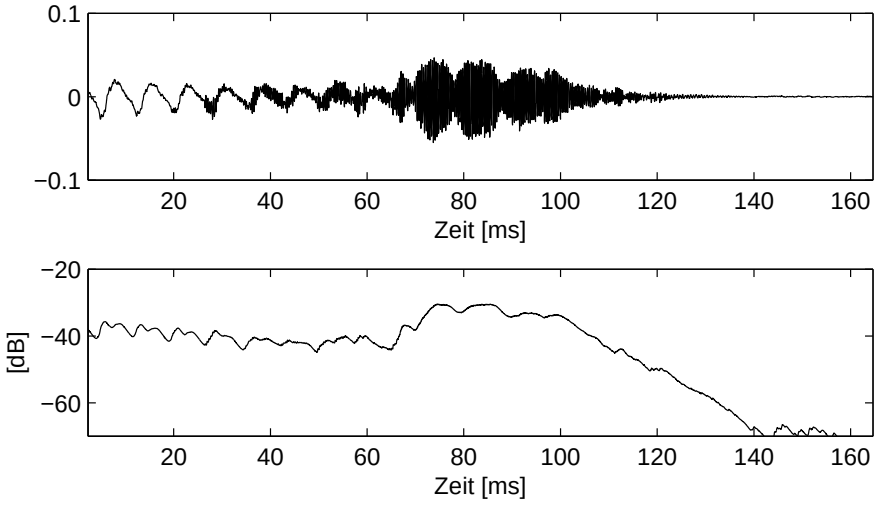
Als weitere elementare Kenngrösse wird ein Mass für die Signalenergie der einzelnen Analyseabschnitte berechnet. Hierfür wird das gleichgerichtete Sprachsignal mit einem Tiefpass gefiltert. Der Tiefpass wurde als IIR-Filter erster Ordnung in der Direktform 1 realisiert.

In Figur 8 ist das Bode-Diagramm des verwendeten Filters dargestellt.

Figur 9 zeigt ein Signalabschnitt eines Sprachsignals und die aus dem Signalbetrag durch Filtern erhaltene Signalenergie. Für die Kurzzeitanalyse wird die Signalenergie in dem jeweiligen Analyseabschnitt gemittelt.



Figur 8: *Bode-Diagramm des Tiefpass Filters*



Figur 9: *Beispiel eines Signalauschnittes und der daraus extrahierte Energieverlauf*

4 Detektion

4.1 Extraktion der Merkmale

Sprachsignale verändern sich fortwährend im Verlaufe der Zeit. Einzelne ihrer Merkmale ändern sich aber langsamer als das Signal selber, sind also über eine gewisse Zeit annähernd stationär. Deshalb genügt für die Messung dieser Merkmale bei der Abtastung eine erheblich geringere Frequenz als für die Abtastung des Signals. Diese Tatsache macht man sich bei der Kurzzeitanalyse zu Nutze. Das Signal wird mithilfe einer Gewichts- oder Fensterfunktion in kleinere Abschnitte unterteilt. Dafür mussten im Wesentlichen drei Parameter festgelegt werden: die Gewichtsfunktion, die Fensterlänge und das Abtastintervall.

Für die Detektion wurde ein Hammingfenster der Länge 500 Samples und ein Abtastintervall von 300 Samples gewählt, was bei einer Abtastrate des Signals von 44.1 kHz ungefähr 11 ms bzw. 7 ms entspricht. Damit wird, ausgehend von den Beobachtungen aus den Spektren, für die Detektion eine genügend hohe Zeit- und Frequenzauflösung erreicht. Aus den so erhaltenen Analyseabschnitten werden nun mit Hilfe der Autokorrelation und der LPC-Analyse die in Kapitel 3 beschriebenen Merkmale extrahiert.

Bei der LPC-Analyse erwies sich eine Prädiktorordnung von 44 als geeignet. Somit könnte im gleichverteilten Fall ein Formant pro kHz erfasst werden unter der allgemeinen Annahme, dass sich ein Formant mit zwei Koeffizienten beschreiben lässt.

In das Mass für die zeitliche Veränderung der Formanten fliessen die Werte der vergangenen fünf Analyseabschnitte mit ein. Dies entspricht 1500 Samples oder einem Zeitraum von ungefähr 35 ms.

4.2 Merkmalsselektion

Die gefundenen Charakteristiken der Störgeräusche lassen eine Eingrenzung der Merkmale zu. So sind relevante Unterschiede bei den Frikativen insbesondere in einem Frequenzbereich zwischen 1.6 und 12 kHz auszumachen. Formanten die ausserhalb dieses Bereiches liegen werden nicht berücksichtigt. Sind keine Übereinstimmungen zwischen den Maxima und den dazugehörigen Frequenzen der AKF und den Formanten des LPC-Spektrums vorhanden, werden die Formanten ebenfalls nicht berücksichtigt. Dabei muss auf die beschränkte Auflösung der AKF insbesondere bei hohen Frequenzen geachtet werden und entsprechend eine gewisse Abweichung bei den Frequenzen toleriert werden. Die Formanten werden weiter eingegrenzt auf diejenigen, deren Amplitude grösser ist als der Mittelwert des LPC-Betragsspektrums. Zuletzt dient noch ein Schwellwert für die Signalenergie als Selektionskriterium. Dieser Schwellwert wurde empirisch für den vorhandenen Signalkorpus auf -50 dB festgelegt.

4.3 Kombination der Merkmale und Gewichtung

Aus den extrahierten und selektierten Merkmalen wird nun für jede Übereinstimmung zwischen Maximum in der AKF und im LPC-Spektrum ein Merkmalsvektor für den Formanten zusammengestellt. Dieser setzt sich aus den elementaren Merkmalen und Kombinationen derselben

folgendermassen zusammen:

$M1$ Die Amplitude des Formanten

$M2$ Die relative Amplitude des Formanten zum Mittelwert des LPC-Betragsspektrums

$M3$ Die logarithmierte Bandbreite des Formanten

$M4$ Die Mittenfrequenz des Formanten

$M5$ Die Standardabweichung für den zeitlichen Verlauf der Formantfrequenz

$M6$ Die Standardabweichung für den zeitlichen Verlauf der Formantamplitude

$M7$ Der Amplitudenunterschied zwischen Formant und dem linken spektralen Minimum (tiefere Frequenz)

$M8$ Der Amplitudenunterschied zwischen Formant und dem rechten spektralen Minimum (höhere Frequenz)

$M9$ Der logarithmierte Frequenzunterschied zwischen Formantfrequenz und Frequenz des linken Minimums

$M10$ Der logarithmierte Frequenzunterschied zwischen Formantfrequenz und Frequenz des rechten Minimums

$M11$ Das Verhältnis zwischen der relativen Amplitude des Formanten und seiner logarithmierten Bandbreite

$M12$ Das Verhältnis zwischen Amplitudenunterschied und logarithmiertem Frequenzunterschied des Formanten und des linken spektralen Minimums

$M13$ Das Verhältnis zwischen Amplitudenunterschied und logarithmiertem Frequenzunterschied des Formanten und des rechten spektralen Minimums

Die Logarithmierung der Bandbreite und der Frequenzunterschiede wurde erst im Verlauf der Optimierungsversuche vorgenommen. Der Grund für diese Anpassung war die Feststellung, dass der hörbare Unterschied zwischen zwei Formanten mit einer Bandbreite von 50 Hz bzw. 100 Hz deutlich höher ist als bei Bandbreiten von 500 und 550 Hz. Eine entsprechende Abhängigkeit konnte auch bei den Frequenzabständen zwischen Formanten und spektralen Minima beobachtet werden.

4.4 Klassifikation durch Bewertung mit linearem Ansatz

Für die Bewertung wurde ein einfaches lineares Modell benutzt. Das Modell verwendet folgenden additiven Ansatz:

$$cx = c_0 \cdot x_0 + c_1 \cdot x_1 + \cdots + c_n \cdot x_n = Y \tag{2}$$

c ist der Merkmalsvektor und x enthält die Gewichte der Merkmale. Daraus ergibt sich Y, welches als Mass für das Vorhandensein eines Störgeräusches dient.

Bei der Umsetzung dieses Modells stellten sich im Wesentlichen die zwei folgenden Probleme:

Die akustische Klassifizierung fand auf Ebene des gesamten Frikativs statt. Es wurde also der Frikativ als Ganzes beurteilt und nicht einzelne Abschnitte innerhalb des Frikativs. Auf der Suche nach den Charakteristiken der Störgeräusche in den Spektrogrammen konnte aber festgestellt werden, dass die typischen Merkmale oft nur zeitlich begrenzt vorhanden sind. Eine direkte Optimierung des Gewichtsvektors konnte deswegen nicht vorgenommen werden, da sowohl Abschnitte mit Störgeräusch als auch ohne vorhanden gewesen wären. Somit konnten daraus keine wertvollen Ergebnisse resultieren.

Desweiteren werden die Merkmalsvektoren pro Formant zusammengestellt. Ein Analyseabschnitt enthält im Allgemeinen aber auch nach der Merkmalselektion (Abschnitt 5.3) mehrere Formanten, die einen unterschiedlich starken Einfluss auf die Wahrnehmbarkeit eines Störgeräusches haben. Es stellte sich dann die Frage, wie sich der Einfluss der einzelnen Merkmalsvektoren auf die Gesamtbeurteilung eines Analyseabschnitts auszuwirken hat. Versuche mit einem additiven wie auch mit einem multiplikativen Ansatz brachten unbefriedigende Ergebnisse. Bessere Resultate konnten erzielt werden, indem pro Analyseabschnitt nur der Merkmalsvektor eines Formanten für die Bewertung verwendet wurde. Ausgewählt wird dafür derjenige mit dem grössten geschätzten Effekt auf die Wahrnehmbarkeit eines Störgeräusches.

Aus den oben genannten Gründen war eine unmittelbare Optimierung des Gewichtsvektors anhand der extrahierten Merkmale und der anfangs erstellten Klassifikation nicht sinnvoll. Eine feinere zeitliche Eingrenzung der Störgeräusche wäre hierfür bereits eine grosse Hilfe. Da die zeitliche Lokalisierung der Geräusche aber akustisch äusserst schwierig ist, hätte diese visuell erfolgen müssen.

Um dies zu umgehen, wurde eine Annäherung bestehend aus zwei Schritten gewählt. In einem ersten Schritt wurde von Hand ein initialer Gewichtsvektor ermittelt, der im Grossen und Ganzen bereits plausible Ergebnisse bei der Detektion liefert. Dabei wurde versucht, die Gewichte so anzupassen, dass eine möglichst gute Übereinstimmung zwischen den Beobachtungen im Spektrogramm und den Bewertungen der einzelnen Analyseabschnitte erzielt wurde. Um das nötige Experimentieren mit den Gewichtsparametern in Grenzen zu halten wurden hierfür allerdings nur 5 Merkmale ($M2, M5, M11, M12, M13$) berücksichtigt. Zu diesem Zweck wurden die kombinierten Merkmale eingeführt. Gestützt auf diese Ergebnisse kann so pro Frikativ der Merkmalsvektor ausgewählt werden, dessen gewichtete Summe am grössten ist. Dies geschieht unter Annahme, dass der ermittelte Maximalwert für den Gesamthöreindruck ausschlaggebend ist. Die Wahl und die Qualität des Initialvektors sind folglich massgebend für die Ergebnisse der Detektion.

Zur späteren Beurteilung der Klassifizierungsleistung wurden die Daten an dieser Stelle per Zufall in ein Trainings- und in ein Testset (265 zu 132) unterteilt. Aus den so erhaltenen Merkmalsvektoren des Trainingssets wurde in einem zweiten Schritt ein neuer Gewichtsvektor ermittelt, bei dem anstatt der kombinierten Merkmale die grundlegenden Merkmale berücksichtigt werden konnten. Hierfür wurde das Y im linearen Modell durch die akustisch ermittelte Klasse der Frikative ersetzt. Als Lösung des Minimierungsproblems

$$\min_x \|cx - Y\|_2 \tag{3}$$

des überbestimmten Gleichungssystems erhält man mit

$$x = (c^t c)^{-1} c^t Y \tag{4}$$

einen neuen Gewichtsvektor x. Diese Vorgehensweise hat den Vorteil, dass die Varianz innerhalb der Klassen minimiert wird, wodurch sich die Unterscheidbarkeit zwischen den Klassen verbessert. Dafür wird in Kauf genommen, dass durch die Optimierung mit diskreten Zielwerten bereits gute Ergebnisse, die eine eindeutige Klassifikation zulassen, unnötigerweise verschlechtert werden. Durch eine Anpassung der Fehlerfunktion, so dass gezielt nur die unklaren Resultate optimiert werden anstatt der allgemeinen Minimierung der Varianz, könnte vermutlich ein besseres Ergebnis erzielt werden. Aus zeitlichen Gründen musste allerdings darauf verzichtet werden. Mit dem neuen Gewichtsvektor wurde dann wiederum eine Bewertung der Frikative vorgenommen.

Für die Klassifizierung wurden pro Wahrnehmungsklasse der Erwartungswert und die Varianz der geschätzten Werte vom Trainingsset berechnet. Unter der Annahme einer Normalverteilung der Daten wurden als Entscheidungsgrenzen die Schnittpunkte der Verteilungen gewählt.

4.5 Resultate der Detektion

4.5.1 Ergebnisse für die gesamte Frikativsammlung

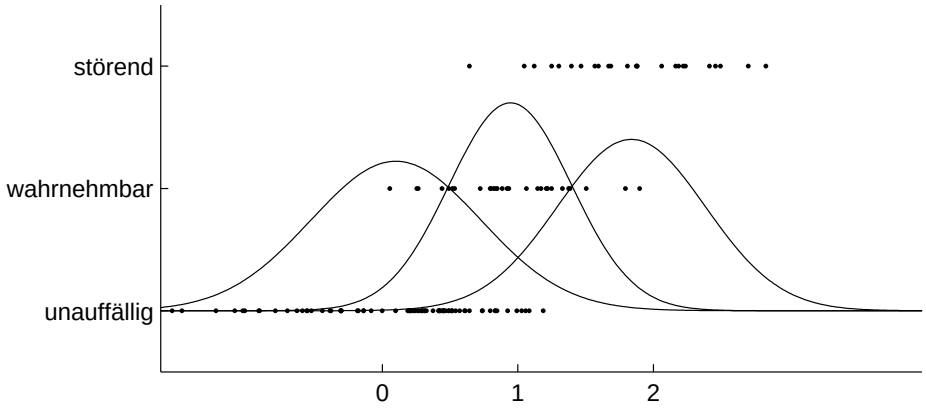
Die Ergebnisse der Klassifikation durch das Detektionsprogramm sind in Form einer Konfusionsmatrix in Tabelle 3 dargestellt. Die erreichte Klassifizierungsrate über das Testset beträgt dabei 73.48%. Je nach Wahl des Trainings- und Testsets variiert die Klassifizierungsleistung aufgrund der ungleichmässigen Verteilung der Klassen um $\pm$5%.

		Akustisch zugeordnete Klasse		
		$K0$	$K1$	$K2$
Ermittelte Klasse	$K0$	57	4	0
	$K1$	22	21	6
	$K2$	0	3	19

Tabelle 3: *Konfusionsmatrix Klassifikation aller Frikative des Testsets*

Am schlechtesten schneidet die Klasse $K1$ mit einem positiven Vorhersagewert (Recall) von 75.00% und einer Sensitivität (Precision) von 42.85% ab. Damit ergibt sich ein F-Wert von 54.54%, welches als Mass für die Gesamtgüte der Klassifikation dient. Es sind inbesondere viele Fehlklassifikationen von Frikativen der Klasse «unauffällig» als «wahrnehmbar» zu verzeichnen.

Da die Schwierigkeit der Zuordnung der Frikative in der Klasse $K1$ auch bereits bei der akustischen Klassifikation bestand, wurde das Modell hier probeweise auf eine Klassifikation zwischen den Klassen «unauffällig» und «störend» beschränkt. In diesem Fall beträgt die Klassifizierungsrate 96.15%. R- und P-Wert betragen 96.00% bzw. 88.89%. Damit ergibt sich ein deutlich höherer F-Wert von 92.31%.



Figur 10: *Schätzwerte und Gaussverteilungen aller Klassen aus dem Testset*

		Akustisch zugeordnete Klasse	
		$K0$	$K2$
Ermittelte Klasse	$K0$	76	1
	$K2$	3	24

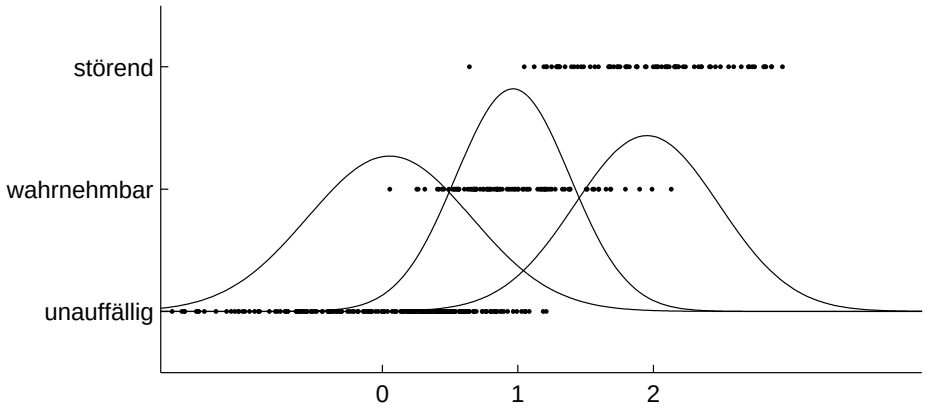
Tabelle 4: *Konfusionsmatrix Klassifikation der $K0$ und $K2$ des Testsets*

Der Vollständigkeit halber seien hier noch die Werte über das gesamte Set aufgeführt. Die zugehörigen Konfusionsmatrizen sind in Tabelle 5 und 6 dargestellt.

Die Klassifizierungsrate bei der Unterscheidung zwischen den drei Klassen beträgt 75.06%. Der minimale F-Wert von 57.87% wird wiederum von der Klasse «wahrnehmbarer» Störgeräusche bestimmt. Bei einer Beschränkung auf die Klassen $K0$ und $K2$ erreicht die Detektion eine Klassifizierungsrate von 97.72% und ein F-Wert von 95.54%.

		Akustisch zugeordnete Klasse		
		$K0$	$K1$	$K2$
Ermittelte Klasse	$K0$	171	8	0
	$K1$	61	68	17
	$K2$	0	13	59

Tabelle 5: *Konfusionsmatrix Klassifikation aller Frikative des gesamten Sets*



Figur 11: *Schätzwerte und Gaussverteilungen aller Klassen aus Test- und Trainingsset zusammen*

		Akustisch zugeordnete Klasse	
		$K0$	$K2$
Ermittelte Klasse	$K0$	226	1
	$K2$	6	75

Tabelle 6: *Konfusionsmatrix Klassifikation der $K0$ und $K2$ des gesamten Sets*

4.5.2 Ergebnisse für die verschiedenen Frikativlaute

Da die absoluten Häufigkeiten einzelner Laute zum Teil ohnehin sehr klein sind, wird an dieser Stelle auf eine Auflistung der Ergebnisse des Testsets alleine verzichtet. Diese sind im Anhang A zu finden. Stattdessen werden die Ergebnisse für die einzelnen Laute über das gesamte Set aufgeführt. Die dazugehörigen Konfusionsmatrizen sind in Tabelle 7 dargestellt.

		Akustisch zugeordnete Klasse				
		$K0$	$K1$	$K2$	$K0$	$K2$
[s]	$K0$	57	7	0	93	0
	$K1$	39	50	11		
	$K2$	0	0	9	3	20
[z]	$K0$	22	0	0	34	1
	$K1$	13	12	5		
	$K2$	0	0	0	1	4
[ʃ]	$K0$	0	0	0	0	0
	$K1$	0	0	0		
	$K2$	0	1	21	0	21
[ʒ]	$K0$	0	0	0	1	0
	$K1$	3	1	1		
	$K2$	0	12	26	2	27
[f]	$K0$	24	1	0	28	0
	$K1$	4	3	0		
	$K2$	0	0	3	0	3
[v]	$K0$	68	0	0	70	0
	$K1$	2	2	0		
	$K2$	0	0	0	0	0

Tabelle 7: *Konfusionsmatrizen der Klassifikation über das gesamte Set nach Lauten geordnet*

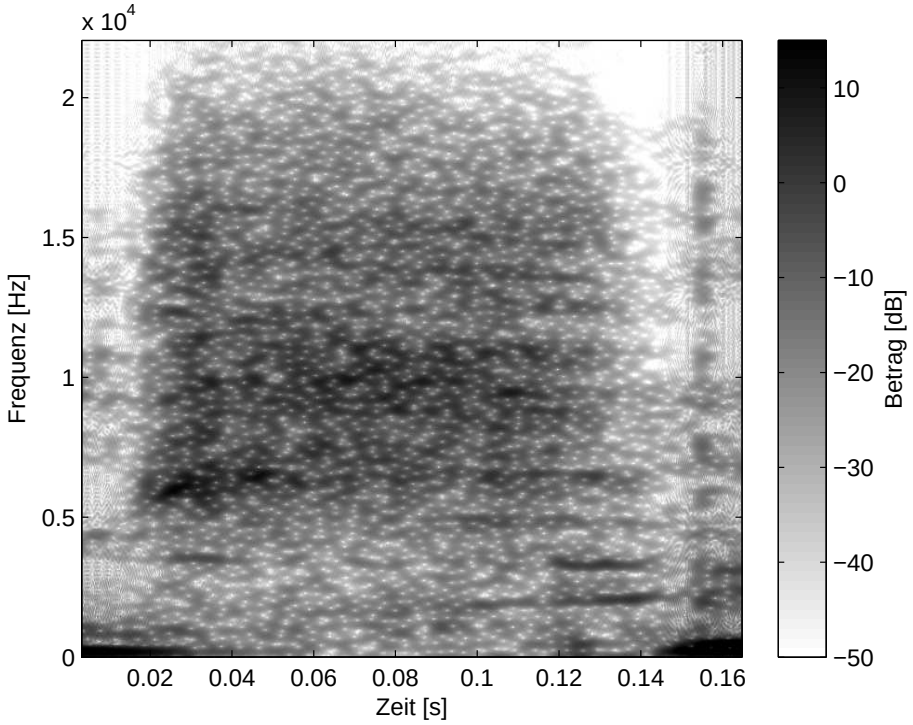
Insgesamt kann man feststellen, dass eine Unterscheidung beschränkt auf die Klassen $K0$ und $K2$ («unauffällig» und «störend») wenige Fehlklassifikationen aufweist. Die niedrigste Klassifikationsrate wird bei den stimmhaften s-Lauten erreicht. Unter Berücksichtigung aller drei Klassen unterscheiden sich die Klassifikationsraten der Laute deutlich voneinander.

Insbesondere bei den s-Lauten zeigt sich eine hohe Fehlklassifikationsrate, so dass keine sinnvolle Unterscheidung zwischen den Klassen «unauffällig» und «wahrnehmbar» und zwischen «wahrnehmbar» und «störrend» möglich ist.

4.6 Schwierigkeiten bei der Detektion

Eine Untersuchung der Gründe für die hohen Fehlklassifikationsraten bei s-Lauten brachte die folgenden Ergebnisse. An Lautübergängen, vor allem von s-Lauten zu Vokalen, findet man bei einem Grossteil der falsch positiv klassifizierten Frikativen zeitlich begrenzte aber deutlich ausgeprägte Formanten im Bereich zwischen 2000 und 5000 Hz und stark periodische Anteile im Signal. In Figur 12 ist ein solches Beispiel dargestellt. Dies sind die typischen Charakteristika

wie sie bei Pfeifgeräuschen zu finden sind. Entsprechend wird hier ein Störgeräusch detektiert. Akustisch ist ein Pfeifen aber an den betroffenen Übergängen kaum bis gar nicht wahrzunehmen, insbesondere nicht beim Anhören eines längeren Signalabschnitts.



Figur 12: *Typisches Beispiel eines falschklassifizierten s-Lautes*

5 Elimination

5.1 Einleitung

Aus der Analyse der Störgeräusche kann man folgern, dass sich die Störgeräusche in erster Linie durch starke periodische Anteile im Signal mit entsprechend ausgeprägten Formanten im Spektrum auszeichnen. Folglich ist anzunehmen, dass eine Dämpfung der entsprechenden Anteile auch die Wahrnehmbarkeit eines Störgeräusches verringert.

Eine direkte Dämpfung des Frequenzbereichs um den Formanten mit Hilfe eines Bandstopp-Filters erwies sich schnell als ungeeignet. Besonders bei Formanten mit asymmetrischen Flanken traten dabei unschöne Nebeneffekte auf. Bei einer zu hohen Bandbreite des Filters verändert sich durch die gleichmässige Dämpfung die Klangfarbe des Lautes merklich. Wird die Bandbreite zu klein gewählt oder sind die Flanken asymmetrisch, wird der bestehende Formant zwar gedämpft, dafür entstehen daneben aber neue spektrale Maxima. Der Störeffekt wird dadurch oft noch verstärkt.

5.2 Aufbau des Eliminationsprogramms

Aus den oben genannten Gründen wurde ein anderer Ansatz verfolgt, basierend auf der Idee eines Begrenzers, mit dem Ziel, eine selektivere Dämpfung der Geräusche zu erreichen.

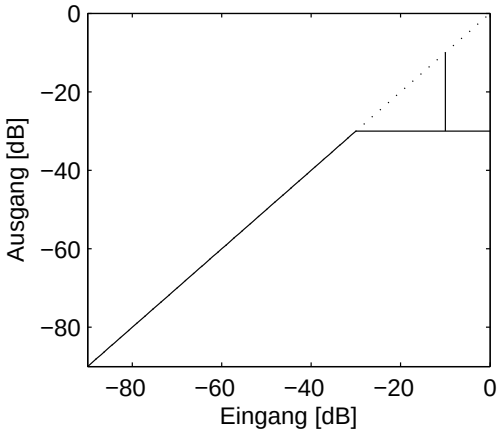
Am Eingang des Eliminationsverfahrens wird das Signal dupliziert. Eine Kopie des Zeitsignals wird in einem ersten Schritt mit einem Bandpass gefiltert. Die Bandmittenfrequenz und die Bandbreite werden entsprechend den Ergebnissen aus der Detektion passend zum Formanten gesetzt der abgeschwächt werden soll. Realisiert wurde das Filter als IIR Filter 2. Ordnung in der Direktform 1. Um den zeitlichen Veränderungen der Formanten gerecht zu werden, werden die Koeffizienten des Filters für jeden Analyseabschnitt neu berechnet. Die exakten Berechnungsformeln für die Filterkoeffizienten sind in [BJ04] zu finden.

Der Betrag des gefilterten Signals wird darauf mit dem in Kapitel 3.4 beschriebenen Tiefpass gefiltert um ein Mass für die im bandpassgefilterten Signal vorhandene Energie zu bekommen. Der Wert der Energie wird in dB umgerechnet.

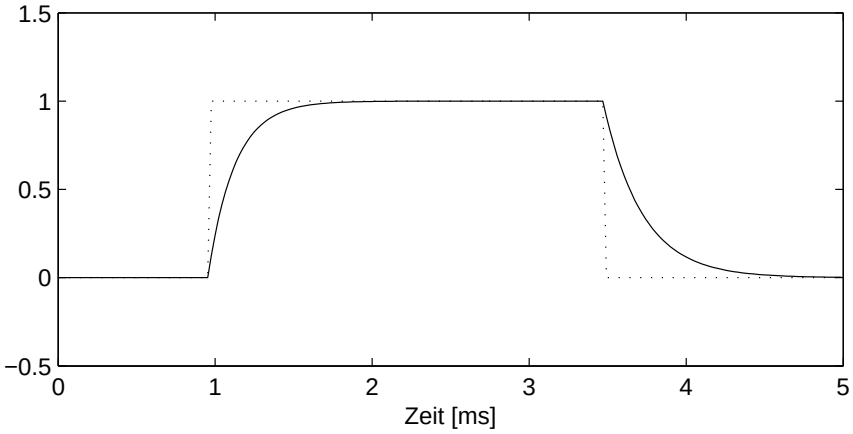
Anhand einer statischen Eingangs-Ausgangs-Kennlinie mit einem vorgegebenen Schwellwert, ab dem der Ausgang begrenzt werden soll, wird berechnet, um wie viel dB das Signal gedämpft werden muss, um der Vorgabe zu entsprechen. Figur 13 zeigt ein Beispiel einer solchen Kennlinie mit einem Schwellwert von -30dB. Bei einem Eingangssignal von -10dB muss eine Dämpfung um 20dB erfolgen. Als Schwellwert für Formanten eines Störgeräusches wurde der Mittelwert der Amplituden der beiden spektralen Minima links und rechts von dem Formanten gewählt.

Aus der Umwandlung von diesem Dämpfungswert in den linearen Bereich ergibt sich ein Faktor g. Das bandpassgefilterte Signal multipliziert mit $(1-g)$ entspricht dem Anteil, der den Schwellwert übersteigt und somit eliminiert werden soll.

Zur Vermeidung hörbarer Regelvorgänge und Verzerrungen aufgrund von kurzzeitigen Änderungen des Skalierungsfaktors g wird die zeitliche Veränderung der Dämpfung zusätzlich träge gemacht wie in Figur 14 dargestellt. Die Zeitkonstanten für die Anstiegs- und die Ab-



Figur 13: *Eingangs-Ausgangs-Kennlinie zur Ermittlung des Dämpfungsfaktors*



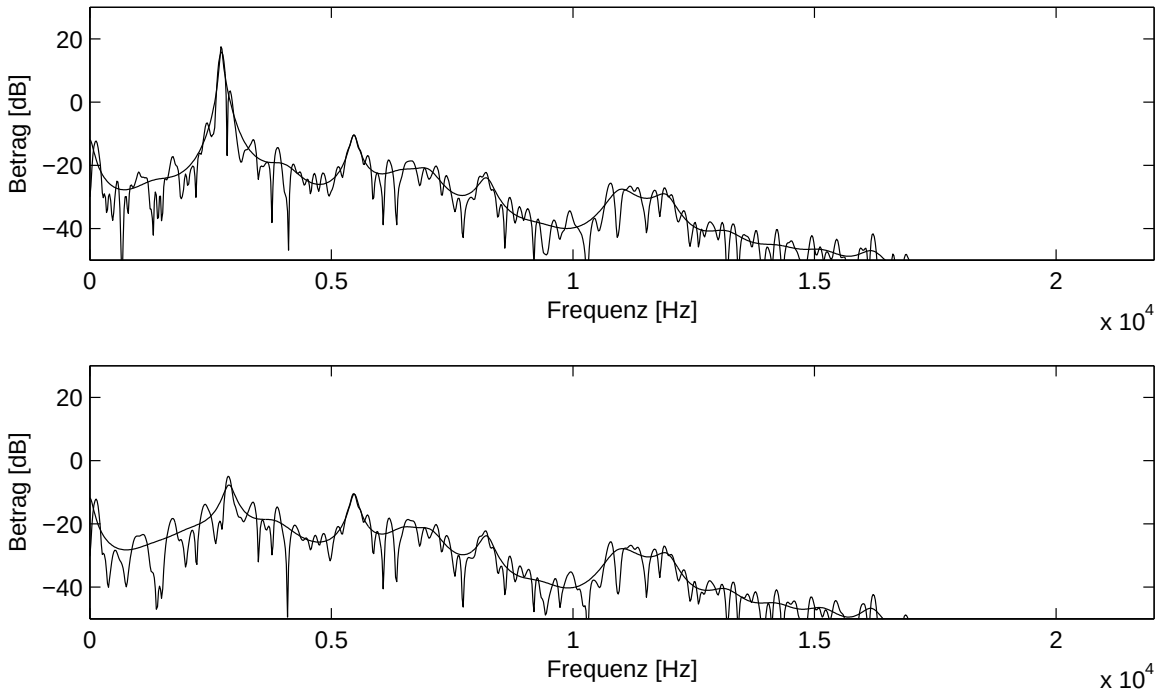
Figur 14: *Sprungantwort des Verzögerungsgliedes für Rechtecksignal(gepunktet) als Eingangssignal*

klingzeit des Verzögerungsgliedes 1. Ordnung wurden empirisch ermittelt. Einerseits darf die Verzögerung nicht zu gross sein, so dass die Dämpfung zu spät beginnt, andererseits soll sie aber auch nicht zu klein sein, da ansonsten jede geringste Überschreitung des Schwellwertes zu einer Dämpfung führt. Damit wird allerdings auch in Kauf genommen, dass der Pegel kurzzeitig den Schwellwert übersteigen kann ehe die Dämpfung einsetzt. In einem letzten Schritt wird von der zweiten Kopie des Zeitsignals am Eingang das bandpassgefilterte Signal multipliziert mit $(1-g)$ subtrahiert und ausgegeben.

5.3 Resultate der Elimination

Zur Beurteilung der Qualität der Elimination und zur Wahl geeigneter Parameter wurden auf der einen Seite die Signale vor und nach der Elimination akustisch miteinander verglichen, auf der anderen Seite wurden die Veränderungen auch optisch verfolgt, da die Charakterisierung der Geräuschtypen ebenso erfolgte.

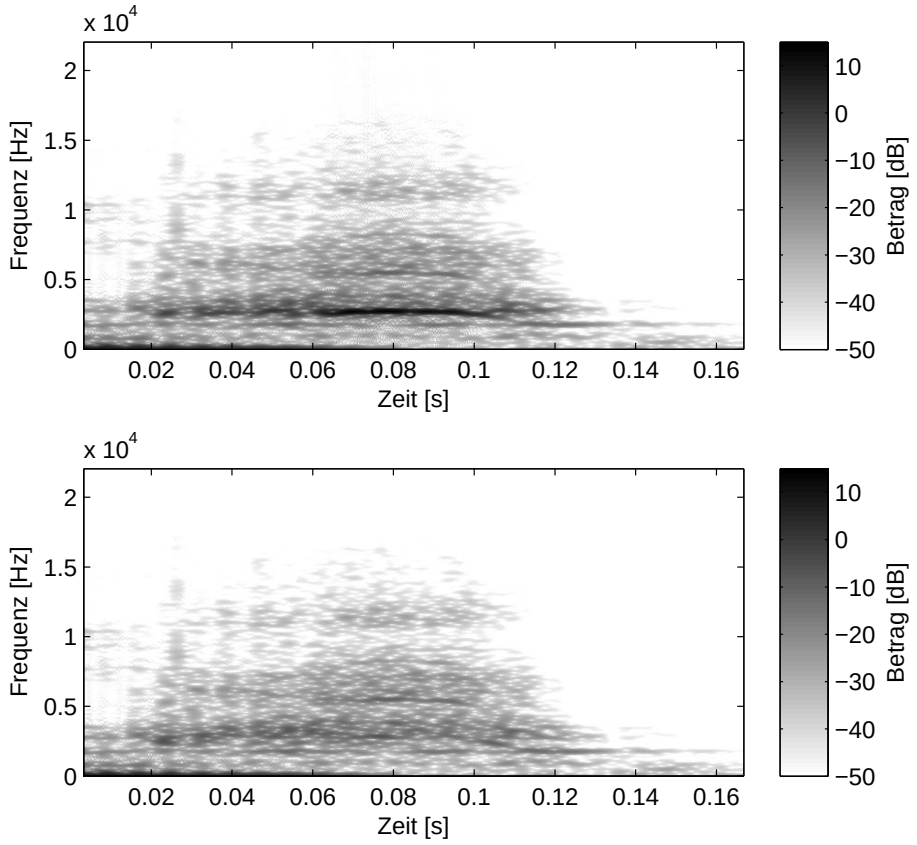
Probleme bei dem verwendeten Eliminationsverfahren stellten sich vor allem im Umgang mit Veränderungen der Parameter. Schnelle und grosse Änderungen der Parameter, insbesondere bei den Bandpasskoeffizienten, führten zu unschönen Verzerrungen und Effekten wie Klicken im Signal. Beim Bandpass musste eine maximal zulässige Verschiebung der Mittenfrequenz pro Zeit festgelegt werden und zusätzlich zwischen den Parametern des alten und des neuen Analyseabschnitts eine Überblendung eingefügt werden. Beim Dämpfungsfaktor brachte ein einfaches Verzögerungsglied eine Verbesserung der resultierenden Klangeigenschaften. Dies zusammen führt zu einem erheblichen Mehraufwand an benötigter Rechenleistung. Die gewonnene Qualität rechtfertig jedoch diesen Schritt.



Figur 15: *LPC-Spektrum vor (oben) und nach (unten) der Elimination eines Pfeifgeräusches*

Bei leichten bis mittelstarken Störgeräuschen lässt sich auf diesem Wege eine deutliche Verringerung der Wahrnehmbarkeit des Geräusches erzielen ohne negative Folgen für den allgemeinen Klangcharakter.

Für Frikative mit erheblichen Störgeräuschen gelingt das Entfernen der Störgeräusche nur bedingt. Die Geräusche werden abgeschwächt, sind aber meist trotzdem noch wahrnehmbar. Durch eine zusätzliche Verstärkung der Dämpfung können in diesen Fällen bessere Ergebnisse erzielt werden. Hierfür wird der Schwellwert für die Dämpfung zusätzlich gesenkt. Setzt man den Schwellwert allerdings zu tief, ändert sich die Klangfarbe des Lautes hörbar.



Figur 16: *Spektrogramm eines Lautes vor (oben) und nach (unten) der Elimination eines Pfeifgeräusches*

In Figur 15 ist von einem Frikativ mit einem Pfeifgeräusch das LPC-Spektrum eines Analyseabschnittes vor und nach dem Elimination dargestellt. Gedämpft wurde dabei der Formant bei 3 kHz. Figur 16 zeigt die dazugehörigen Schmallbandspektrogramme des Frikativs.

6 Diskussion und Ausblick

Die wohl anspruchsvollste und zugleich auch wichtigste Aufgabe dieser Arbeit bestand in der auditiven Klassifikation und Beurteilung der vorhandenen Signale. Da es sich dabei nur um eine qualitativ erfassbare Grösse handelt ist das Ergebnis natürlich stark subjektiv geprägt. Insbesondere musste ich zu meinem Erstaunen feststellen, dass die Einteilung der Frikative nicht nur von persönlichen Faktoren beeinflusst wird, sondern auch situative Faktoren eine Rolle spielen. So kam es im Verlaufe der Arbeit ein paar Mal vor, dass ich das Gefühl hatte, die anfangs vorgenommene Zuordnung der Frikative sei nicht immer richtig gewesen und ich mich zu dem Zeitpunkt anders entscheiden würde. Angesichts dieser Tatsache sind die stark unterschiedlichen Klassifikationsraten, je nachdem ob alle Frikative berücksichtigt werden oder nur die zweifelsfrei klassifizierbaren, besser verständlich.

Das Detektionsverfahren liefert bei Berücksichtigung aller drei Klassen eine Gesamtklassifizierungsrate von 75.06%. Bei einer Beschränkung auf die Klassen «unauffällig» und «störend» wird eine Klassifizierungsrate von 97.72% erreicht. Der verwendete Ansatz mit Merkmalen aus der Autokorrelation, dem LPC-Spektrum und Informationen über deren zeitlichen Verlauf erwies sich als brauchbare Basis für die Detektion der Störgeräusche. Die Unterscheidungsstärke reicht jedoch nicht für eine eindeutige automatische Klassifikation der Frikative.

Für eine Verbesserung der Detektionsleistung sind sowohl eine genauere zeitliche Lokalisation der Störgeräusche als auch eine feinere Unterteilung der Geräuschstärken bei den Frikativen von Nöten. Diese Aufgabe ist äusserst zeitaufwändig und verlangt ein sehr feines Gehör, würde dafür aber den Einsatz auch komplexerer Modelle bei der Detektion erlauben.

Eine interessante Untersuchung hierzu wäre weiter, ob eine repräsentativere Klassifizierung der Frikative durch mehr Personen ein klareres Ergebnis bringen würde, oder ob die stark subjektiven Höreindrücke eher das Gegenteil bewirken und ein Verschmieren der Resultate zur Folge haben würden.

Eine Schwierigkeit beim Übertragen und bei der Anpassung des Detektionsverfahrens auf andere Stimmen dürften der Mangel an Trainingsbeispielen sein. Die relativ grosse Anzahl an Störgeräuschen im verwendeten Signalkorpus ist eher als Ausnahme zu betrachten bei Stimmen, die als Grundlage einer Sprachsynthese dienen sollten.

Das vorgestellte Eliminationsverfahren kann die Wahrnehmbarkeit der Störgeräusche klar verringern. Es ist jedoch auf eine genaue Detektion dieser angewiesen. Bei Frikativen mit sehr starken Störgeräuschen gelingt eine vollständige Entfernung dieser allerdings nicht.

Literatur

[BJ04] R. Bristow-Johnson. Cookbook formulae for audio eq biquad filter coefficients. 2004. http://www.musicdsp.org/files/Audio-EQ-Cookbook.txt.

[PK08] B. Pfister and T. Kaufmann. *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. Springer Verlag, 2008.

A Klassifikationsergebnisse des Testsets nach Lauten geordnet

		Akustisch zugeordnete Klasse				
		$K0$	$K1$	$K2$	$K0$	$K2$
[s]	$K0$	20	3	0	33	0
	$K1$	15	14	2		
	$K2$	0	0	2	2	4
[z]	$K0$	9	0	0	14	1
	$K1$	6	4	3		
	$K2$	0	0	0	1	2
[ʃ]	$K0$	0	0	0	0	0
	$K1$	0	0	0		
	$K2$	0	1	5	0	5
[ʒ]	$K0$	0	0	0	1	0
	$K1$	1	0	1		
	$K2$	0	2	10	0	11
[f]	$K0$	5	1	0	5	0
	$K1$	0	2	0		
	$K2$	0	0	2	0	2
[v]	$K0$	23	0	0	23	0
	$K1$	0	1	0		
	$K2$	0	0	0	0	0

Tabelle 8: *Konfusionsmatrizen der Klassifikation des Testsets nach Lauten geordnet*

B Aufgabenstellung

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

Herbstsemester 2008
(SA-2008-30)

Aufgabenstellung

für

Herrn Simon Simonet

Betreuer: T. Ewender, ETZ D97.7
S. Hoffmann, ETZ D97.5

Ausgabe: 22. September 2008
Abgabe: 19. Dezember 2008

Detektion und Elimination von Störgeräuschen bei Frikativlauten

Einleitung

Selbst bei Sprachaufnahmen professioneller Sprecher kommt es vor, dass Frikativlaute (vor allem die Laute [s] und [f]) Pfeif- und Zischgeräusche enthalten. Diese Störgeräusche stellen sowohl bei der Unit Selection als auch bei der Diphonsynthese ein Problem dar.

Bei der Unit Selection werden Sprachsignale synthetisiert, indem aus einem relativ grossen Sprachsignalkorpus möglichst grosse Ausschnitte gesucht und zum gewünschten Sprachsignal zusammengesetzt werden. Dabei fallen Störgeräusche bei immer wiederkehrender Verwendung der gleichen Units auf. Bei der Diphonsynthese werden bei der Synthese am Signal prosodische Veränderungen vorgenommen, die diese Störgeräusche zusätzlich verstärken.

Es wäre deshalb sehr nützlich, ein Verfahren zu haben, mit welchem diese Störgeräusche detektiert werden könnten. Je nach Art und Genauigkeit des Detektionsverfahrens könnte dieses dann unterschiedlich eingesetzt werden. Man könnte beispielsweise bereits die laufenden Sprachaufnahmen überwachen (On-line-Einsatz), um festzustellen, ob Störgeräusche in den Frikativen vorhanden sind und sodann eventuell eine Wiederholung des gesprochenen Satzes veranlassen. Auch im Off-line-Einsatz, also bei der Suche nach geeigneten Segmenten in der Sprachsynthese, wäre das Detektionsverfahrens nützlich. So

könnten beispielsweise aus einer umfangreichen Menge von Sprachsignalen die Segmente so ausgewählt werden, dass problematische Frikative vermieden werden.

Diese beiden Anwendungsszenarien unterscheiden sich massgeblich in der bereits verfügbaren Information über die Signale. Im On-line-Verfahren ist keinerlei Information darüber vorhanden, wo im Signal welche Laute sind, da die Verarbeitung in Echtzeit stattfindet und möglicherweise noch keine Trainingsdaten für die Segmentierung zur Verfügung stehen. Im Off-line-Verfahren kann man jedoch von einer bereits vorhandenen Lautsegmentierung ausgehen, die eine Einschränkung der Verarbeitung auf die problematischen Signalabschnitte der Frikative ermöglicht.

Ein weiterer Schritt wäre die Elimination der Pfeif- und Zischgeräusche bei bereits vollständig aufgenommenen und vorhandenen Signalen. Bei der Erstellung von Stimmen für verschiedene Sprachen sind die Sprecher nur für den kurzen Zeitraum der Aufnahmen vor Ort und stehen insbesondere später nicht mehr zur Verfügung. Es wäre deshalb wünschenswert, bei bereits durchgeführten Aufnahmen die vorhandenen Störgeräusche zu eliminieren.

Problemstellung

Das Ziel dieser Semesterarbeit ist, zu untersuchen, mit welchen Methoden Pfeif- und Zischgeräusche in Frikativen erkannt werden können. Da es sich um eine Arbeit mit starkem Fokus auf Signalanalyse handelt, werden verschiedene Analysemethoden der Signalverarbeitung zum Einsatz kommen. Ein zweites Ziel ist, die detektierten Signalabschnitte so zu verändern, dass die Störgeräusche akustisch nicht mehr wahrnehmbar sind, selbst wenn prosodische Veränderungen am Signal vorgenommen werden.

Vorgehen

Aufgrund der Literatur gibt es bisher kaum einschlägige Untersuchungen, welche für diese Arbeit wegweisend sein könnten. Es wird deshalb das folgende Vorgehen empfohlen:

1. Erstellen Sie eine Sammlung von Frikativbeispielen aus dem vorhandenen Signalkorpus. Für einen ersten Ansatz kann von einer vorhandenen automatischen Lautsegmentierung der Signale ausgegangen werden, die eventuell noch manuell verfeinert werden müsste. Diese Sammlung soll Fälle enthalten, bei denen problematische Störgeräusche sowohl wahrnehmbar als auch nicht wahrnehmbar sind. Vorteilhaft wäre eine Klassifikation anhand der Art des Störgeräusches und der Stärke der Wahrnehmbarkeit.

2. Analysieren Sie die Beispiele unter Verwendung verschiedener Merkmale. Untersuchen Sie sodann, welche davon zur Detektion der Störgeräusche brauchbar sind und verwendet werden können. Möglicherweise nützliche Merkmale sind: spektrale Merkmale, Formantdetektionsmethoden [1], Autokorrelationsmethode, Spectral Flatness Measure.

Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht zu dokumentieren (siehe dazu [2]), der in gedruckter und in elektronischer Form (als PDF-Datei) 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] H. Hanson, et al. A system for finding speech formants and modulations via energy separation. *IEEE Transactions on Speech and Audio Processing*, 2(3):436–443, July 1994.

[2] B. Pfister. *Richtlinien für das Verfassen des Berichtes zu einer Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/richtlinien_bericht.pdf).

[3] B. Pfister. *Hinweise für die Präsentation der Semester- oder Diplomarbeit.* Institut TIK, ETH Zürich, März 2004.
(http://www.tik.ee.ethz.ch/~spr/SADA/hinweise_praesentation.pdf).

[4] B. Pfister and T. Kaufmann. *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung.* Springer Verlag (ISBN: 978-3-540-75909-6), 2008. http://www.springer.com/978-3-540-75909-6.

Zürich, den 25. September 2008