Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Sprache mit Tierstimmencharakteristik

Samuel Groner und Andreas Scherrer

Semesterarbeit SA-2005-04

Wintersemester 2004/05

Institut für Technische Informatik
und Kommunikationsnetze

Betreuer: Harald Romsdorfer

Verantwortlicher: Prof. Dr. L. Thiele

Inhaltsverzeichnis

Zusammenfassung **4**

1 Einleitung **4**

2 Gepresste Stimme **5**
2.1 Quellmaterial . . . 5
2.2 Analyse . . . 5
2.2.1 Ungeglättetes Spektrogramm . . . 5
2.2.2 Geglättetes Spektrogramm . . . 5
2.2.3 Statistisches Spektrum . . . 7
2.2.4 Konstruktionsversuch . . . 8
2.2.5 Verbesserung des Konstruktionsversuches . . . 9
2.2.6 Dekonstruktionsversuch . . . 10
2.2.7 Übertragen des Effektes auf andere Sprecher . . . 10
2.3 Resultate . . . 11
2.4 Weitere Versuche . . . 11
2.4.1 Verschiedene Filter verwenden . . . 11
2.4.2 Vokaltrakt modifizieren . . . 11
2.4.3 Formanten parametrisch analysieren . . . 12
2.4.4 Formantenverlauf analysieren . . . 12

3 Krächzstimme **13**
3.1 Quellmaterial . . . 13
3.2 Analyse . . . 13
3.2.1 Erste Versuche . . . 13
3.2.2 Statistisches Spektrogramm . . . 14
3.2.3 Cepstrogramm . . . 14
3.2.4 Betrachtung kurzer Signalausschnitte . . . 14
3.2.5 Betrachtung des Fehlersignals . . . 15
3.2.6 LPC-Analyse/Synthese als Dekonstruktionsversuch . . . 16
3.2.7 LPC-Analyse/Synthese als Konstruktionsversuch . . . 18
3.3 Resultate . . . 20
3.4 Weitere Versuche . . . 20
3.4.1 Impulsstärke korrigieren . . . 21
3.4.2 Systematik der Anregung analysieren . . . 21
3.4.3 Weitere Merkmale berücksichtigen . . . 21

A Matlab-Funktionen **22**
A.1 Funktionen im Verzeichnis `basic_functions/` . . . 22
A.2 Funktionen im Verzeichnis `basic_functions/utilities/` . . . 24
A.3 Funktionen im Verzeichnis `basic_functions/old/` . . . 25
A.4 Skripts im Verzeichnis `bericht/bilder/` . . . 25

A.5 Skripts im Verzeichnis `bericht/hoerbeispiele/` 25
A.6 Diskussion der Funktion `STAT_ANALYSE06` 25
A.7 Diskussion der Funktion `LPC_ANALYSESYNTHESE_FO03` 26

B Beschreibung der beiliegenden CD **28**

C Aufgabenstellung **29**

Literaturverzeichnis **31**

Zusammenfassung

Für diese Arbeit beschäftigten wir uns mit Sprachsignalen, die von verstellten Stimmen verschiedener Sprecher stammen. Wir versuchten herauszufinden, welche charakteristischen Merkmale die Stimmen zweier Charaktere eines Kasperli-Theaters von normal gesprochenem Text unterscheiden. Namentlich handelt es sich hierbei um einen Riesen und einen Raben bzw. Papagei.

Das Ergebnis für den Riesen fiel dabei etwas weniger befriedigend aus als das des Papageien, für beide Fragestellungen fanden wir aber Ansätze. Beim Riesen identifizierten wir das Anheben des zweiten Formanten, beim Papageien eine unperiodische Anregung als für das Charakteristische der Stimme mit- oder sogar hauptverantwortliche Merkmale.

Unsere Untersuchungen könnten für beide Effekte weitergeführt werden. Ideen hierfür haben wir jeweils am Ende des entsprechenden Kapitels notiert. Wir konnten diese Ideen aus Zeitgründen nicht weiterverfolgen.

1 Einleitung

Ein menschlicher Sprecher hat die Möglichkeit, seine Stimme sehr verschieden klingen zu lassen. Menschen sind in der Lage, so zu sprechen, dass ein Zuhörer das Gefühl hat, der Sprecher imitiere gerade einen Papagei oder eine Schlange. Oder aber er spricht «ganz normal», also so, wie er oder sie in einer ganz alltäglichen Situation spricht.

In unserer Arbeit ging es darum, herauszufinden, wie sich das aufgezeichnete Sprachsignal (technisch gesehen) verändert, wenn ein Sprecher seine Stimme verstellt.

Sämtliche analysierten Signale stammen aus Aufnahmen von Kasperli-Theatern, aus der Sammlung der Sprachverarbeitungsgruppe an der ETH Zürich oder sind selbst aufgezeichnet. Die Kasperli-Hörspiele werden von drei Personen, einer Frau und zwei Männern, gesprochen. Ausgewählt haben wir diese Hörspiele, weil sie viele verschiedene Charakteren enthalten und jeder Sprecher, da sie nur zu dritt sind, mehrere Rollen spricht. Dies ermöglichte uns einen aussagekräftigen Vergleich zwischen verstellter und unverstellter Stimme.

Die Sprachsignale der Sprachverarbeitungsgruppe und die selbst aufgenommenen Signale benutzten wir hauptsächlich für Konstruktionsversuche. Wir haben sie verwendet, weil sie leicht zugänglich waren, bzw. der Text frei gewählt werden kann.

2 Gepresste Stimme

Im ersten Teil unserer Semesterarbeit beschäftigten wir uns mit Sprachsignalen eines Sprechers, der einen Riesen imitiert. Die unverstellte Stimme des Sprechers klingt eher tief; um den Riesen zu imitieren, verstellt er sie aber zusätzlich noch. Der Effekt, den dieses Verstellen hat, lässt sich mit «gepresst» beschreiben. Weiter wirkt die Stimme in einem gewissen Masse aggressiv.

Da es sich bei einem Riesen um ein Phantasiewesen handelt, gibt es keinen allgemeingültigen Begriff davon, wie ein Riese zu tönen hat. Es scheint aber vernünftig oder verbreitet, mit dem Riesen aufgrund seiner Grösse eine tiefe, mächtige Stimme zu verbinden. Trotzdem ist es nicht einfach, objektiv zu entscheiden, ob ein Sprachsignal mehr oder weniger als ein anderes nach Riese tönt.

2.1 Quellmaterial

Als Datengrundlage für die Analyse dienten uns Sätze und Wörter aus zwei Kasperli-Hörspielen. Der Charakter «Mösli» repräsentiert dabei die unverstellte, der Charakter «Riese» die verstellte Sprechweise des Sprechers. Zwei Hörbeispiele befinden sich auf der beiliegenden CD (Nr. 2.1 (Mösli) bzw. Nr. 2.2 (Riese)). Wir haben hauptsächlich mit den Vokalen [a], [u] und [e] gearbeitet. Da die Signale von einer Audio-CD stammen, beträgt die Samplefrequenz 44.1 kHz. Ein klarer Nachteil dieser Signale war, dass verständlicherweise keine Labelfiles vorhanden waren; einen beträchtlichen Anteil der uns zur Verfügung stehenden Zeit mussten wir daher in das Erstellen von Labelfiles investieren.

Für die Konstruktionsversuche haben wir Signale der Sprachverarbeitungsgruppe und zu den Originalsignalen inhaltlich identische, selbst aufgenommene Sprachsignale hinzugenommen.

2.2 Analyse

Wir haben die vorliegenden Signale nach verschiedenen Kriterien analysiert. Auf die Methoden und Ergebnisse gehen wir nun im Einzelnen ein.

2.2.1 Ungeglättetes Spektrogramm

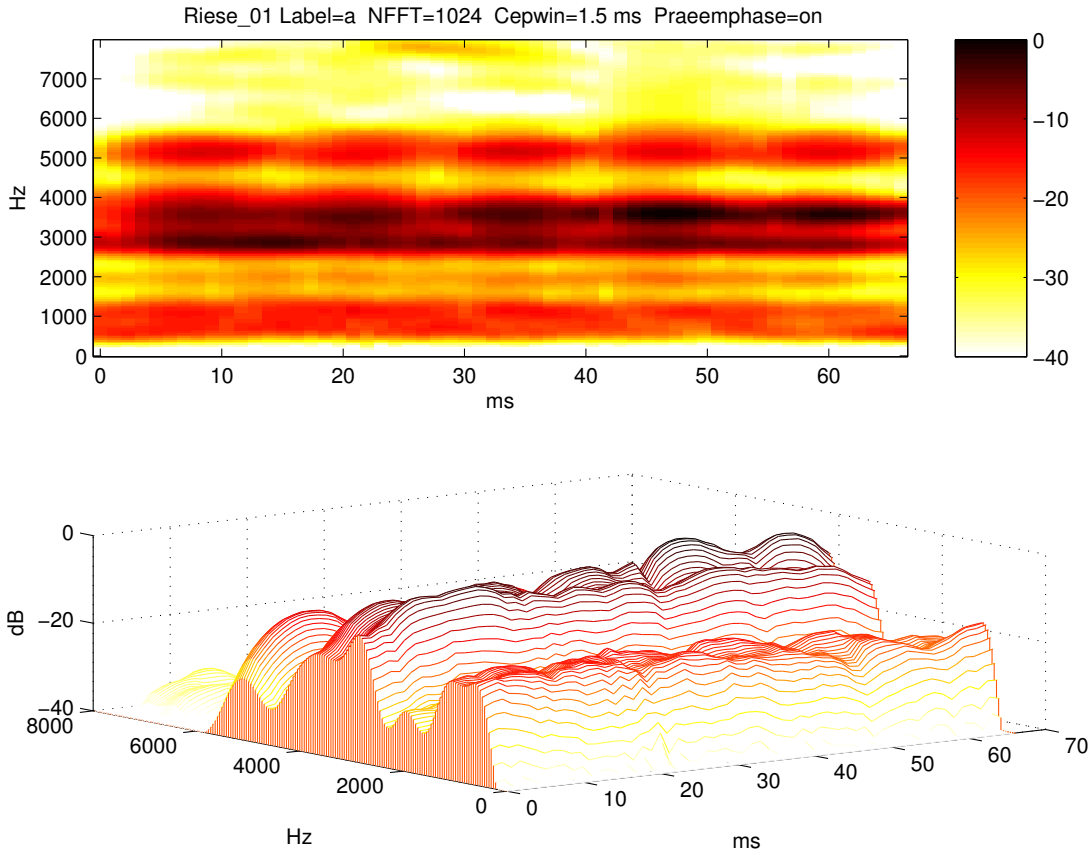
Wir haben die Spektrogramme der ganzen Signale sowie von Ausschnitten, die einzelne Laute repräsentieren, jeweils mit und ohne Präemphase betrachtet. Alle diese Spektrogramme waren nicht geglättetet. In beiden Darstellungen konnten wir, auch mit Variation der zur Verfügung stehenden Parameter (insb. der Fenstergrösse), keine systematischen Auffälligkeiten entdecken.

2.2.2 Geglättetes Spektrogramm

Anschliessend verwendeten wir für die Spektrogramme eine cepstrale Glättung, um die Formantstruktur besser sichtbar zu machen. Für die Glättung verwendeten wir typischer-

weise ein cepstrales Fenster der Länge 1.5 ms. Die detaillierte Implementation ist in der Matlab-Funktion `PLOT_FORMANTS04` einsehbar.

Dargestellt haben wir die geglätteten Spektrogramme einerseits als Colormap, andererseits als Wasserfalldiagramm (siehe Figur 2.1). In der zweitgenannten Darstellungsart betrachteten wir nur Spektrogramme von Einzellauten.



Figur 2.1: *Spektrogramm des Lautes [a], gesprochen mit gepresster Stimme. Sowohl in der Colormap-, wie in der Wasserfalldiagrammdarstellung sind die später gefundenen Resultate (Überhöhung des zweiten Formanten) nur undeutlich ersichtlich.*

Es fiel uns schwer, die (dreidimensional drehbaren) Wasserfalldiagramme auf eine Art zu platzieren, in der man tatsächlich Auffälligkeiten erkennen kann. In einzelnen Fällen gewannen wir eine gute Übersicht über den Verlauf einzelner Frequenzen, eine allgemeingültige Auffälligkeit konnten wir aber in beiden Darstellungsarten nicht feststellen. Insofern blieb der Nutzen der aufwendigeren Darstellung im Wasserfalldiagramm sehr beschränkt.

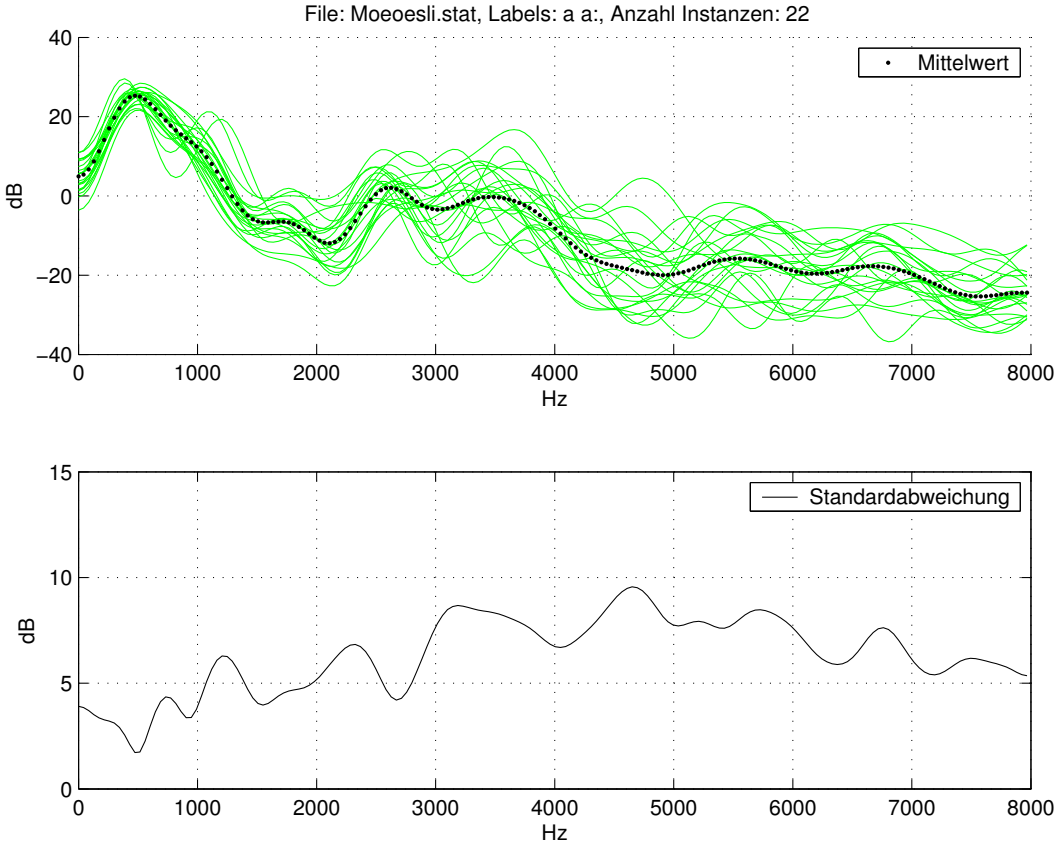
Die später gefundenen Resultate – insbesondere die Überhöhung des zweiten Formanten – sind mit der entsprechenden Kenntnis prinzipiell sichtbar, jedoch nicht von sich aus augenfällig.

2.2.3 Statistisches Spektrum

Um die Signale in einen systematischen Zusammenhang zu bringen, berechneten wir eine statistische Analyse, indem wir den Mittelwert der geglätteten Spektren verschiedener Instanzen gleicher bzw. ähnlicher Laute bildeten.

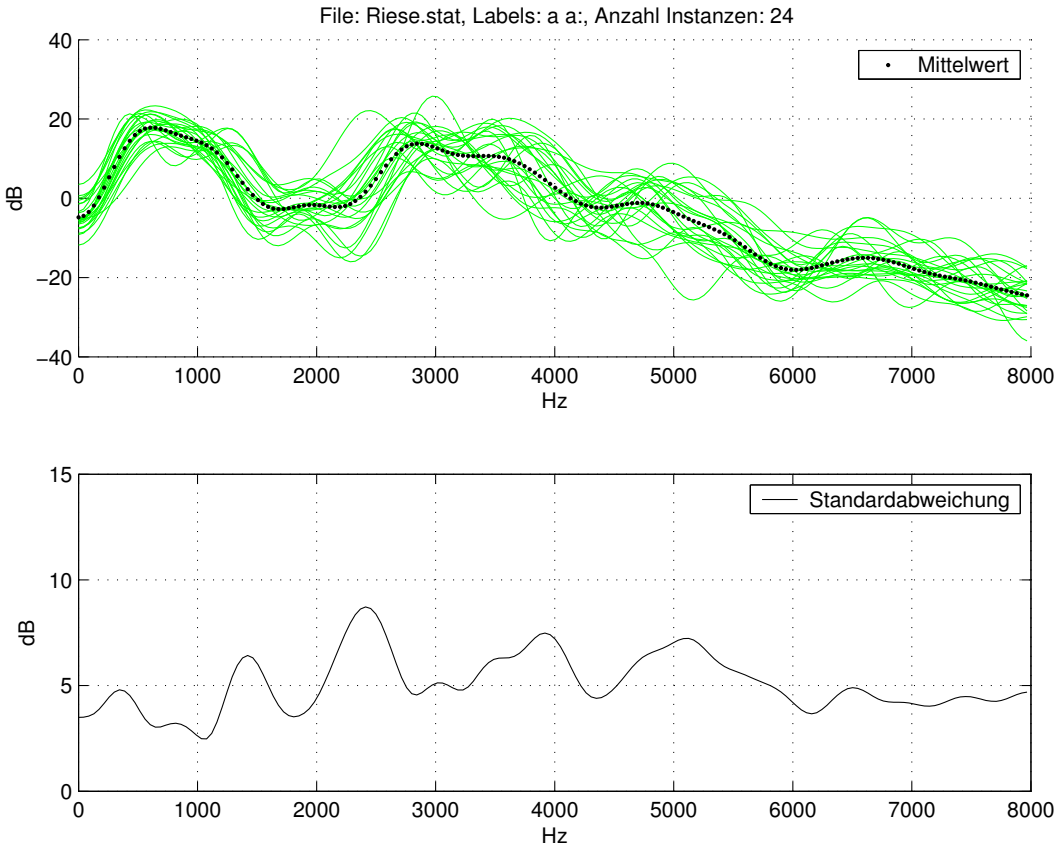
Um abschätzen zu können, wie zuverlässig, bzw. zufällig dieser Mittelwert ist, trugen wir zusätzlich die Varianz der Amplitude pro Frequenz ab. (Siehe Figur 2.2). Wie daraus ersichtlich ist, stimmen die Signale, auch wenn man viele Instanzen verwendet, v.a. in den Bereichen der Formanten gut überein und die Ergebnisse sind dementsprechend als relevant einzustufen. Die spektralen Unterschiede, die sich in den einzelnen Instanzen der Laute ergeben, betreffen insbesondere die Position und nicht die Amplitude der jeweiligen Formanten.

Für eine detaillierte Diskussion der verwendeten Matlab-Funktion `STAT_ANALYSE06` verweisen wir auf Anhang A.6.



Figur 2.2: *Aus 22 Instanzen der Laute [a] und [a:] gewonnenes statistisches Spektrum der unverstellen Sprecherstimme. Die Spektren der einzelnen Lautinstanzen sind grün eingezeichnet.*

Dieser Ansatz förderte etwas zu Tage. Die Signale der Pressstimme haben fast identische Amplitude im ersten und im zweiten Formanten (wobei es sich beim zweiten, je nach



Figur 2.3: *Aus 24 Instanzen der Laute [a] und [a:] gewonnenes statistisches Spektrum einer gepressten Stimme. Im Vergleich zur Figur 2.2 ist klar ersichtlich, wie sich die Formantenstruktur verändert. Besonders auffällig ist die Verstärkung im zweiten Formanten um ungefähr 20 dB. Diese Auffälligkeit ist auch in anderen Vokalen klar ersichtlich.*

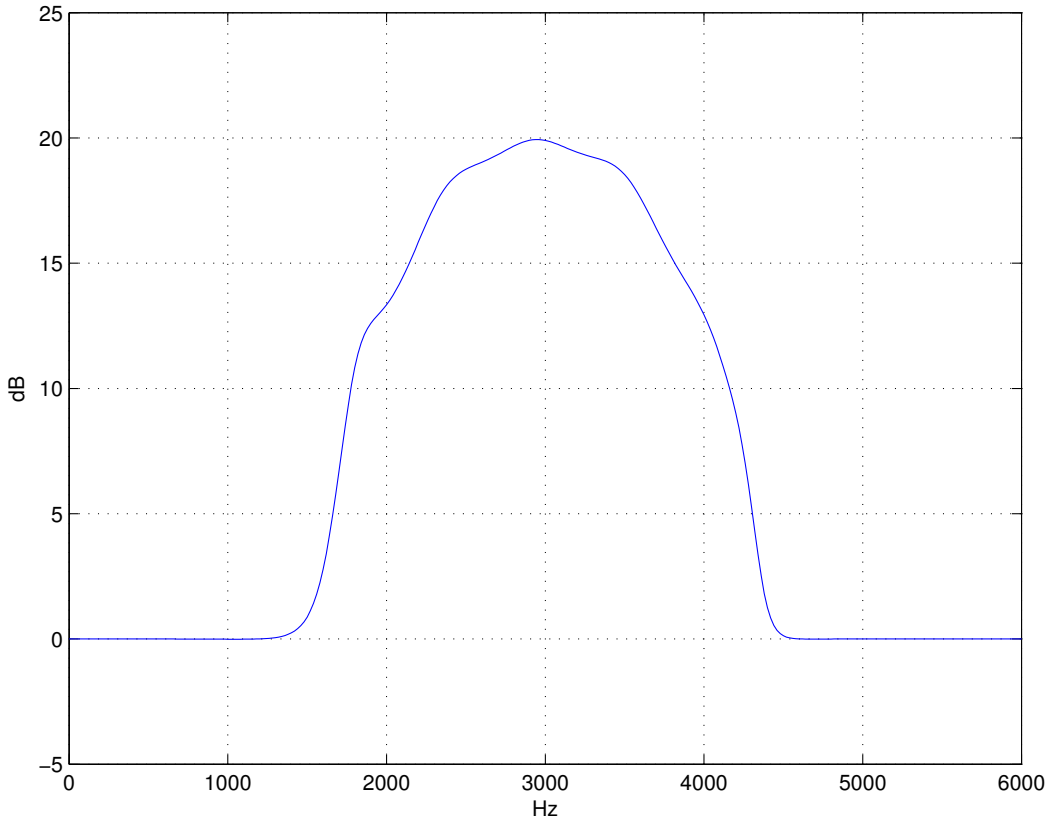
Laut, auch um einen Doppelformanten handeln kann). Die Signale von normal gesprochenen Worten bzw. Lauten, weisen im zweiten gegenüber dem ersten Formanten hingegen einen Amplitudenabfall von ca. 20 dB auf. Figur 2.2 zeigt die statistische Analyse eines normal gesprochenen Lautes, Figur 2.3 die eines gepresst gesprochenen. Wie schon erwähnt, stammen beide Sprachsignale vom gleichen Sprecher.

Wie ein im folgenden beschriebener einfacher Konstruktionsversuch zeigte, scheint diese Beobachtung *eines der Elemente* zu beschreiben, welche die Gepresstheit ausmachen. Wie wir später sehen werden, sind viele andere Merkmale der Sprechweise des Charakters Riese durch diesen Effekt noch nicht hinreichend beschrieben.

2.2.4 Konstruktionsversuch

Zu Beginn haben wir ein normal gesprochenes Wort des Sprechers, welcher auch den Riesen imitiert, mit einem einfachen Filter $\mathcal{F}_a$, das den gesamten Frequenzbereich um den zweiten Formanten um etwa 20 dB anhebt, gefiltert. Schon diese einfache Manipulati-

on verleiht der Stimme eine gewisse Gepresstheit. (Hörbeispiel 2.3). Der entsprechende Frequenzgang des Filters $\mathcal{F}_a$ ist in Figur 2.4 dargestellt.



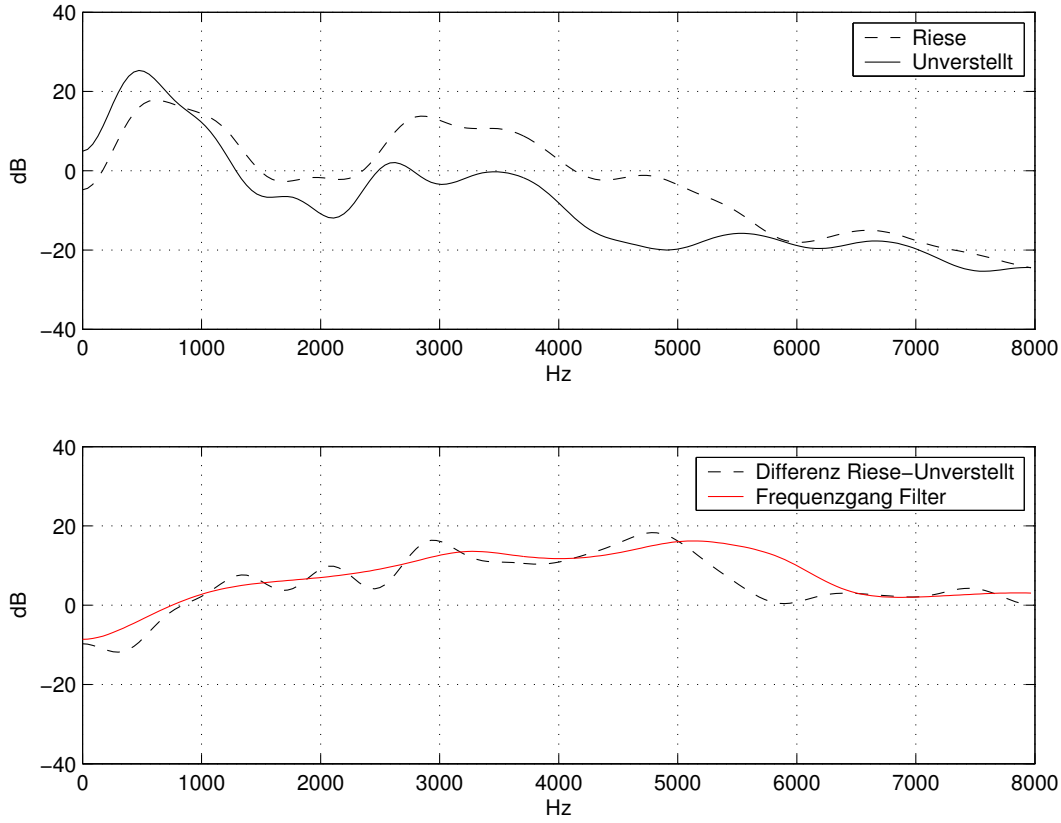
Figur 2.4: *Frequenzgang des Filters, das wir für den ersten Konstruktionsversuch verwendet haben.*

2.2.5 Verbesserung des Konstruktionsversuches

Im folgenden verfeinerten wir das Filter, indem wir den Frequenzgang genauer vorgaben und Filter bis zur Ordnung 12 verwendeten. Um den Frequenzgang, den wir vorgeben wollten, zu erhalten, verwendeten wir wiederum die vorher beschriebenen gemittelten Spektren. Er entstand, indem wir die Differenz des durchschnittlichen Spektrums eines Lautes des Riesen und dem normal gesprochenen (gleichen) Laut desselben Sprechers berechneten (siehe Figur 2.5). Mit diesem Frequenzgang haben wir dann ein neues Filter, welches quasi auf den Sprecher zugeschnitten ist, berechnet. Die beiden dazu verwendeten Matlab-Funktionen `FREQ_DIFF01` und `FILTERBOX03` sind auf der beiliegenden CD verfügbar und im Anhang kurz beschrieben.

Wir bezeichnen dieses über die Differenz der Signale von ein und derselben Person berechnete Filter im weiteren mit $\mathcal{F}_b$.

Filtert man ein unverstellt gesprochenes Signal des *gleichen* Sprechers mit diesem Filter $\mathcal{F}_b$, tritt der gewünschte Effekt deutlicher auf als mit einem Filter $\mathcal{F}_a$. (Hörbei-



Figur 2.5: *Der obere Plot zeigt das statistische Spektrum des Lautes [a], gesprochen mit verstellter und unverstellter Stimme. Der untere Plot zeigt die Differenz dieser Spektren und den Frequenzgang des entsprechenden Konstruktionsfilters $\mathcal{F}_b$.*

spiel 2.4). Filtert man hingegen ein Sprachsignal eines *anderen* Sprechers, funktioniert dieses Vorgehen nicht so gut.

2.2.6 Dekonstruktionsversuch

Uns interessierte an dieser Stelle ebenfalls ein Dekonstruktionsversuch; dieser war durch ein Filter mit zu $\mathcal{F}_b$ inversem Frequenzgang einfach durchzuführen. Wie Hörbeispiel 2.5 zeigt, führte dies zwar zu einer starken Verdumpfung des Klanges, ein auch nur annähernd natürlich klingendes Sprachsignal war aus einer gepressten Stimme auf diese Weise jedoch nicht zu gewinnen.

Dies bestätigte die schon vorher angesprochene Erkenntnis, das das Phänomen der gepressten Stimme so noch keineswegs vollständig beschrieben ist.

2.2.7 Übertragen des Effektes auf andere Sprecher

Weitere zwei Versuche widmeten wir dem Übertragen des Effektes auf die Stimme eines anderen Sprechers. Das Problem hierbei, so schien uns, liegt darin, dass die Formanten bei der Stimme eines anderen Sprechers im Spektrum an anderen Stellen liegen. Deshalb

hebt das über die Differenz der Sinale eines Sprechers ermittelte Filter nicht die richtigen Frequenzen im Signal. Um dies zu korrigieren, berechneten wir die Differenz anstatt aus zwei Lauten ein und desselben Sprechers aus der des zu verändernden mit dem des Riesen. Diese Art Filter bezeichnen wir im weiteren mit $\mathcal{F}_c$. Auch dieser Methode war mässiger Erfolg beschieden. Der Effekt ist zu einem gewissen Grad hörbar, aber weniger ausgeprägt als erhofft (Hörbeispiel 2.6).

Der letzte Versuch bestand darin, die Pole und/oder Nullstellen eines Filters $\mathcal{F}_b$ oder $\mathcal{F}_c$ so zu verschieben, dass sie auf die Formanten der Stimme des zweiten Sprechers passen. Auch so wird der Effekt nicht so gut übertragen wie mit einem Filter $\mathcal{F}_b$ auf dieselbe Stimme. Wir verzichten auf ein Hörbeispiel.

2.3 Resultate

Wir konnten die Formantstruktur in der Riesenstimme als eines von mehreren Merkmalen, die den Charakter der Stimme ausmachen, identifizieren. Konkret heisst das, dass die Frequenzanteile um den ersten bzw. zweiten Formanten im Signal annähernd gleich gross sind. Bei einem normal gesprochenen Signal ist der zweite Formant ca. 20 dB schwächer als der erste.

Dies alleine macht allerdings mitnichten den ganzen Charakter des Riesen aus. Viele Nuancen fehlen in einem konstruierten gegenüber einem originalen Signal. Erschwerend kommt bei der Bestimmung, was genau diese Nuancen sind und wie man sie simulieren könnte, hinzu, dass ein Riese kein real existierendes Wesen ist.

2.4 Weitere Versuche

Folgende weiterführende Versuche könnten angestellt werden. Wir haben sie aus Zeitgründen nicht oder nur ansatzweise durchführen können.

2.4.1 Verschiedene Filter verwenden

Wir verwendeten in allen Versuchen *ein* Filter, mit welchem wir das ganze Signal (ein Wort oder sogar einen Satz) filterten. Der Berechnung des Frequenzgangs des Filters legten wir aber die Differenz eines bestimmten Lautes zugrunde. Aus diesem Grund passt der Filter für genau den bei der Berechnung verwendeten Laut besser als für andere Laute. Der Effekt tritt also für einen Laut stärker auf als für die anderen.

Eine Verbesserungsmöglichkeit besteht sicher darin, für verschiedene Laute verschiedene Filter zu verwenden. Dies setzt die Kenntnis der Lautpositionen im Sprachsignal sowie ein adaptives Filter voraus.

2.4.2 Vokaltrakt modifizieren

Durch einfache Exerimente konnten wir feststellen, dass bei gewissen (typischerweise dünn klingenden) Stimmen eine durch Dauermodifikation und Resamplen hervorgerufene Vo-

kaltraktvergrösserung den subjektiven Eindruck einer gepressten Stimme erhöht; systematische Nachforschungen sind jedoch für eine präzisere Aussage diesbezüglich unerlässlich.

2.4.3 Formanten parametrisch analysieren

Um den Effekt der gepressten Stimme einfacher auf andere Sprecher übertragen zu können, wird es hilfreich sein, die Formantstruktur zu parametrisieren. Passende Parameter scheinen Frequenz, Amplitude und Bandbreite zu sein. Insbesondere der Vergleich dieser Parameter in verstellter und unverstellter Stimme desselben Sprechers dürfte neue Hinweise auf das korrekte Transformieren unverstellter Sprechstimmen geben.

2.4.4 Formantenverlauf analysieren

Für den gehörmässigen Klangfarbeneindruck ist nicht nur das gemittelte Spektrum, sondern auch insbesondere der spektrale Verlauf des Sprachsignales ausschlaggebend. Daher könnte das Untersuchen des zeitlichen Verhaltens der Formantparameter präzisere Resultate ergeben.

3 Krächzstimme

Als Beispiel für eine krächzende Stimme haben wir Signalausschnitte von einer Papagei- und Raben-Imitation beigezogen – beide von einem männlichen Sprecher gesprochen. Vor allem in den Passagen des Raben besteht ein nicht zu vernachlässigender Teil des Textes nur aus «Kraa, kraa», diesen haben wir bei der Analyse ausgeblendet, da er mit natürlichsprachlichen Signalen nichts zu tun hat. Es bleibt genug Material, welches echte Sätze beinhaltet, die sehr gekrächzt gesprochen werden. Hörbeispiele 3.1 und 3.2 für Rabe, bzw. Papagei.

3.1 Quellmaterial

Die gekrächzten Signale stammen von zwei Kasperli-Hörspielen. Zusätzlich haben wir für Konstruktionsversuche ein selbst aufgenommenes Signal verwendet. Dieses ist vom Inhalt her zu einem Sprachsignal aus den Hörspielen identisch.

3.2 Analyse

Wiederum haben wir die Signale auf verschiedene Arten, auf die wir im Folgenden eingehen werden, untersucht.

3.2.1 Erste Versuche

Zu Beginn betrachteten wir das Spektrogramm der Gesamtsignale sowie das einzelner Laute. Auffallend an den längeren Signalabschnitten ist, dass die subjektive und die in den Spektrogrammen ersichtlich Grundfrequenz sehr konstant ist. Offensichtlich macht dies alleine keine krächzende Stimme aus, es ist aber eine mehr oder weniger stark ausgeprägte Gemeinsamkeit der von uns untersuchten Aufnahmen. Weiter fiel auf, dass die Signale relativ grosse Anteile im Bereich von 2 kHz aufweisen.

Wir vermuteten intuitiv, dass der Krächzanteil des Signals bei hohen Frequenzen zu finden sei. Um dies messtechnisch nachvollziehen zu können, verglichen wir die Energie des hochpassgefilterten Signals mit der des tiefpassgefilterten. Obwohl wir mit den Filterparametern experimentierten, konnten wir im Vergleich zu normal gesprochenen Sprachsignalen keine systematische Abweichung der Frequenzbalance feststellen.

Wiederum unter der Annahme, dass das Krächzen des Sprachsignales im hochfrequenten Signalanteil zu finden sei, führten wir einen ersten einfachen Dekonstruktionsversuch durch, indem wir das Sprachsignal tiefpassfilterten. Erstaunlicherweise war das Krächzen selbst mit Filtern hoher Ordnung und niedriger Ansatzfrequenz (1 kHz) nicht zu beseitigen. Unsere erste Annahme scheint damit widerlegt zu sein.

Das Krächzen erinnerte uns bis zu einem gewissen Grad an das Verzerrungsverhalten eines Audioverstärkers; als einfachen Konstruktionsversuch verzerrten wir daher unverstellt gesprochene Signale mit einer nichtlinearen Übertragungsfunktion unter Verwendung von Polynomen unterschiedlicher Ordnung. Durch das Anhören der Resultate konn-

ten wir jedoch, unabhängig von der Verzerrungsstärke, keine klare Ähnlichkeit mit einem Krächzen feststellen.

Eine weitere Ursache des Krächzens vermuteten wir im Phasengang; da die entsprechenden Methoden jedoch komplex in ihrer Entwicklung sind und wir in der Zwischenzeit erfolgsversprechendere Ansätze gefunden hatten, verfolgten wir diese Idee nicht weiter.

3.2.2 Statistisches Spektrogramm

Da wir bei der Analyse der gepressten Stimme mit einer statistischen Spektralanalyse einen Erfolg verbuchen konnten, wendeten wir dieselbe Methode auch auf gekrächzt tönende Stimmen an. Zu erkennen war wiederum eine Überhöhung um 2 kHz. Diese hat, wie ein einfacher Konstruktionsversuch mit einem entsprechenden Filter zeigte, aber offenbar nur wenig mit dem krächzenden Charakter des Signales zu tun. Leider gewannen wir aus dieser Analyse keine weiteren Erkenntnisse.

3.2.3 Cepstrogramm

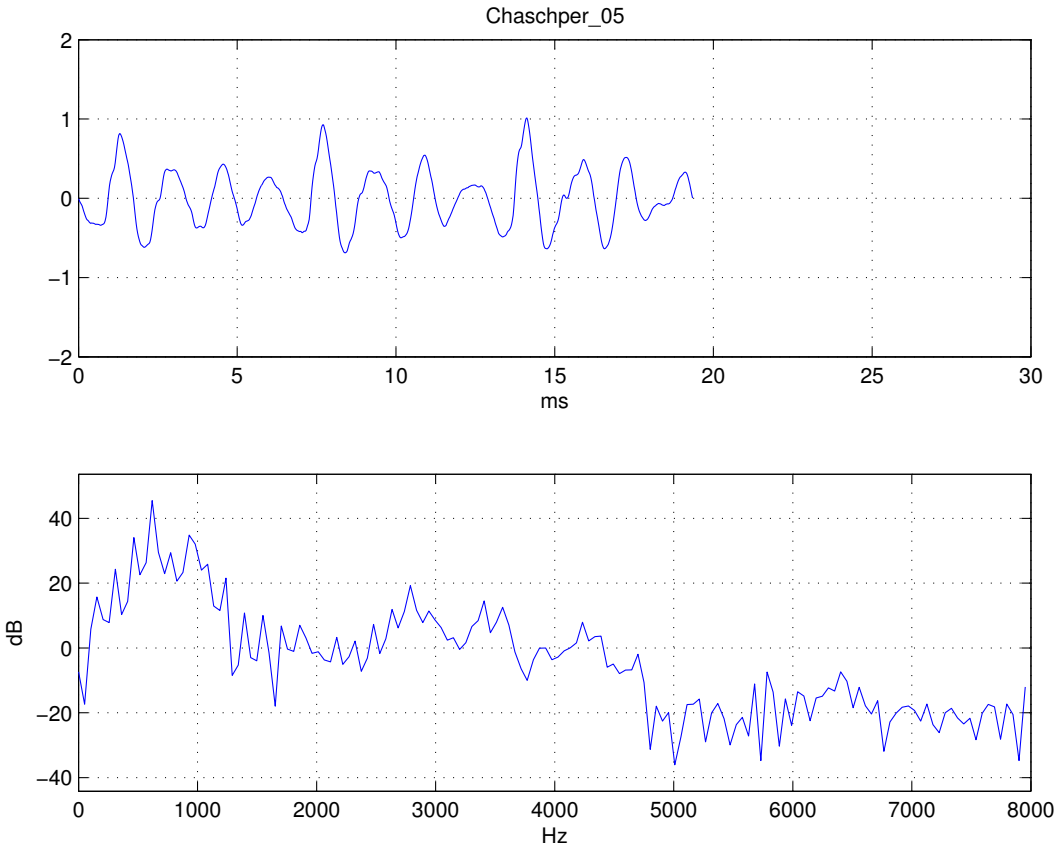
Motiviert durch die Tatsache, dass die Grundfrequenz in den Spektrogrammen so konstant aussah, betrachteten wir das Cepstrogramm. Wir hatten aber keine konkreten Vorstellungen, wonach wir suchten. Das Cepstrogramm lieferte dann auch keine neuen Einsichten in die Thematik.

3.2.4 Betrachtung von kurzen Signalausschnitten im Zeit- und Frequenzbereich

Der gehörmässigen Beurteilung der gekrächzten Signalauschnitte nach zu schliessen, handelt es sich um einen stark nichtlinearen Effekt und nicht nur um eine lineare Frequenzgangänderung. Um diesbezüglich eine genauere Aussage treffen zu können, betrachteten wir Signalauschnitte sowohl von gekrächzt als auch von normal gesprochenen Vokalen; die Länge der Signalauschnitte wählten wir so, dass sie genau drei Perioden entsprach (siehe Figur 3.1 und 3.2). Dazu entwickelten wir die Matlab-Funktion `PLOT_SIGNAL_AND_SPEC01`.

Sehr auffallend in der Frequenzbereichsdarstellung ist die deutliche Überhöhung bei 2 kHz bei allen gekrächzt gesprochenen Signalauschnitten. Dies konnten wir schon früher festellen. Ebenfalls sofort ersichtlich ist die tendenziell komplexere Wellenform im Zeitbereich dieser Signale. Dies lässt sich natürlich durch die unterschiedlichen Spektren der Signale erklären.

Aufgrund der Zeit- und Frequenzbereichsdarstellung des Signalausschnittes in Figur 3.2 würde man die Grundfrequenz auf ungefähr 100 Hz schätzen. Tatsächlich liegt die gehörmässige Empfindung jedoch ungefähr eine Oktave höher. An dieser Stelle ist anzumerken, dass herkömmliche Verfahren zur Grundfrequenzdetektion (z.Bsp. Autokorrelations- oder cepstrale Verfahren) bei diesen Signalen nur sehr bedingt zuverlässige Resultate liefern. Ebenso versagen die meisten Stimmhaft-/Stimmlos-Entscheide.

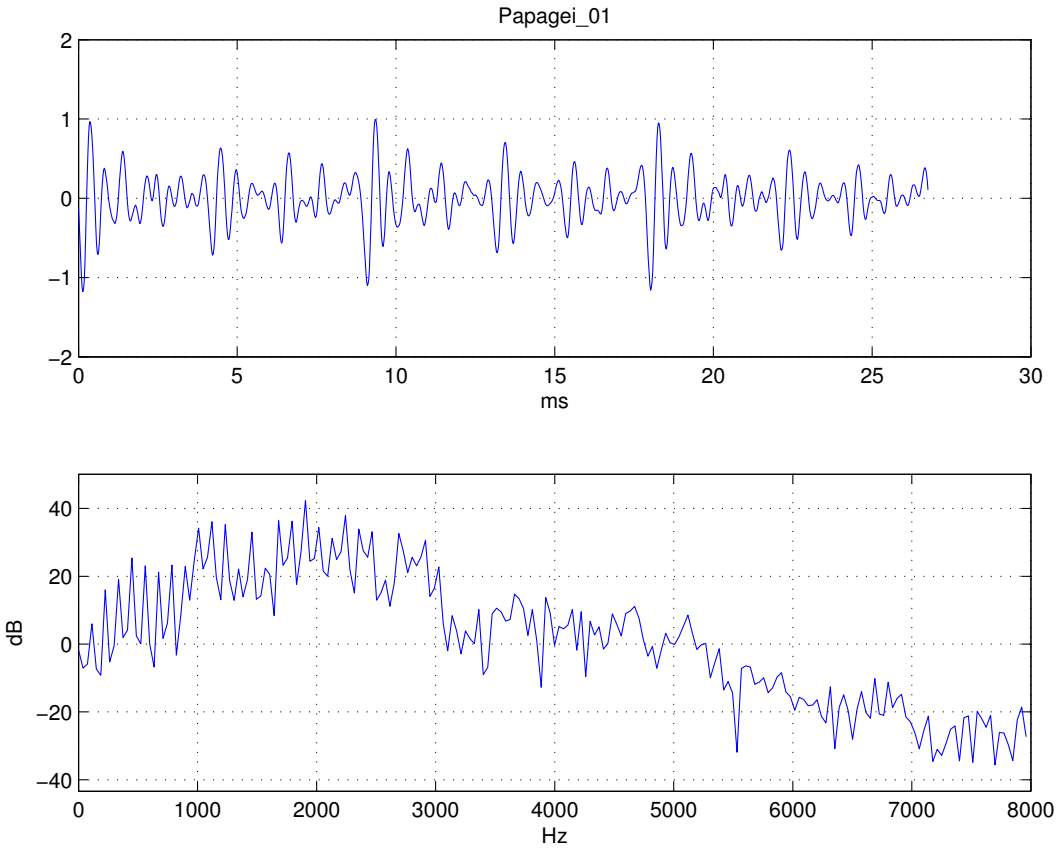


Figur 3.1: *Darstellungen eines auf genau drei Periodenlängen zugeschnittenen Sprachsignales im Zeit- und Frequenzbereich. Es handelt sich dabei um einen Signalausschnitt des unverstellt gesprochenen Lautes [a].*

Bei genauerer Betrachtung der Zeitbereichsdarstellung (insbesondere auch von längeren Signalabschnitten) fiel uns auf, dass die Signale kaum mehr als quasistationär zu bezeichnen sind; in offenbar mehr oder weniger willkürlicher Folge sind zwischen den Impulsen der tieffrequentesten Periodizität weitere Peaks ersichtlich. Intuitiv betrachtet lässt sich dies gut durch ein Ausbrechen der Stimmbänder aus dem periodischen Schwingungsmuster erklären – offenbar Resultat einer stark forcierten Stimme. Aufgrund dieser Einsicht ist das Versagen der genannten Grundfrequenzdetektionsverfahren erklärbar, da diese auf einer Periodizitätsannahme basieren.

3.2.5 Betrachtung des Fehlersignals

Durch die Betrachtung des Fehlersignals aus der LPC-Analyse erhofften wir uns das Erkennen der aperiodischen Impulsfolge der Anregung; tatsächlich ist dies zu einem gewissen Grad ersichtlich, sowohl mit als auch ohne Präemphase (Figur 3.3).

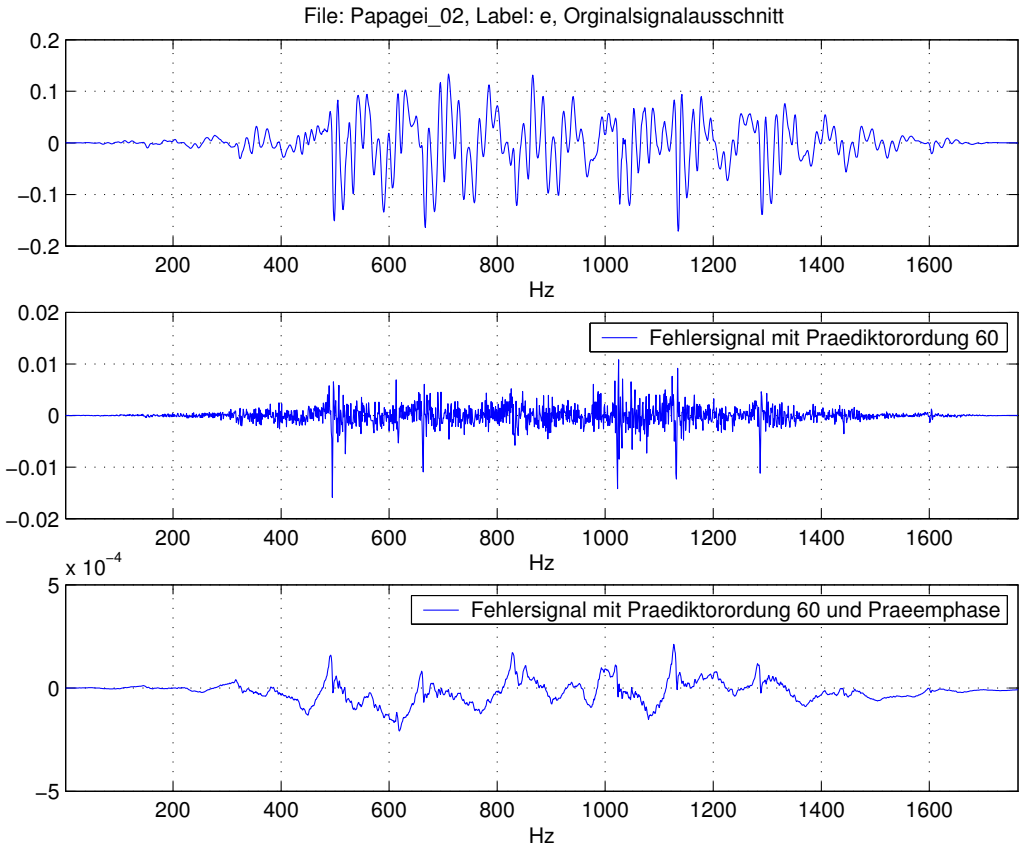


Figur 3.2: *Im Vergleich zu Figur 3.1 ist bei dieser Darstellung die deutliche Anhebung um 2 kHz und die komplexere Wellenform klar ersichtlich – beides mehr oder weniger systematische Merkmale gekrächzter Sprachsignale. Zumindest ansatzweise ist zudem die aperiodische Anregung erkennbar.*

3.2.6 LPC-Analyse/Synthese als Dekonstruktionsversuch

Um weitere Erkenntnisse betreffend dem Phänomen Krächzstimme zu gewinnen, entschieden wir uns für einen Dekonstruktionsversuch. Eine LPC-Analyse/Synthese schien uns geeignet, da bei der Synthese stimmhafte Laute streng mit einem periodischen Signal angeregt werden und dadurch das Krächzen verschwinden sollte. Erschwerend wirkte sich bei der LPC-Analyse der Umstand der schlechten Grundfrequenzdetektion aus; wird die Synthese mit einem unnatürlichen Grundfrequenzverlauf ausgeführt, wirkt das rekonstruierte Signal sofort künstlich und eine Beurteilung anderer Effekte wird durch die entstehende Maskierung praktisch verunmöglicht.

Durch Adaption der in der Sprachverarbeitungsgruppe vorhandenen Grundfrequenzdetektionssoftware liess sich ein mehr oder weniger realistischer Grundfrequenzverlauf erzeugen; durch Beschneiden und Glätten konnten wir das Resultat weiter verbessern. Der resultierende Grundfrequenzverlauf sowie der zugehörige Stimmhaft-/Stimmlos-Entscheid ist in Figur 3.4 dargestellt.

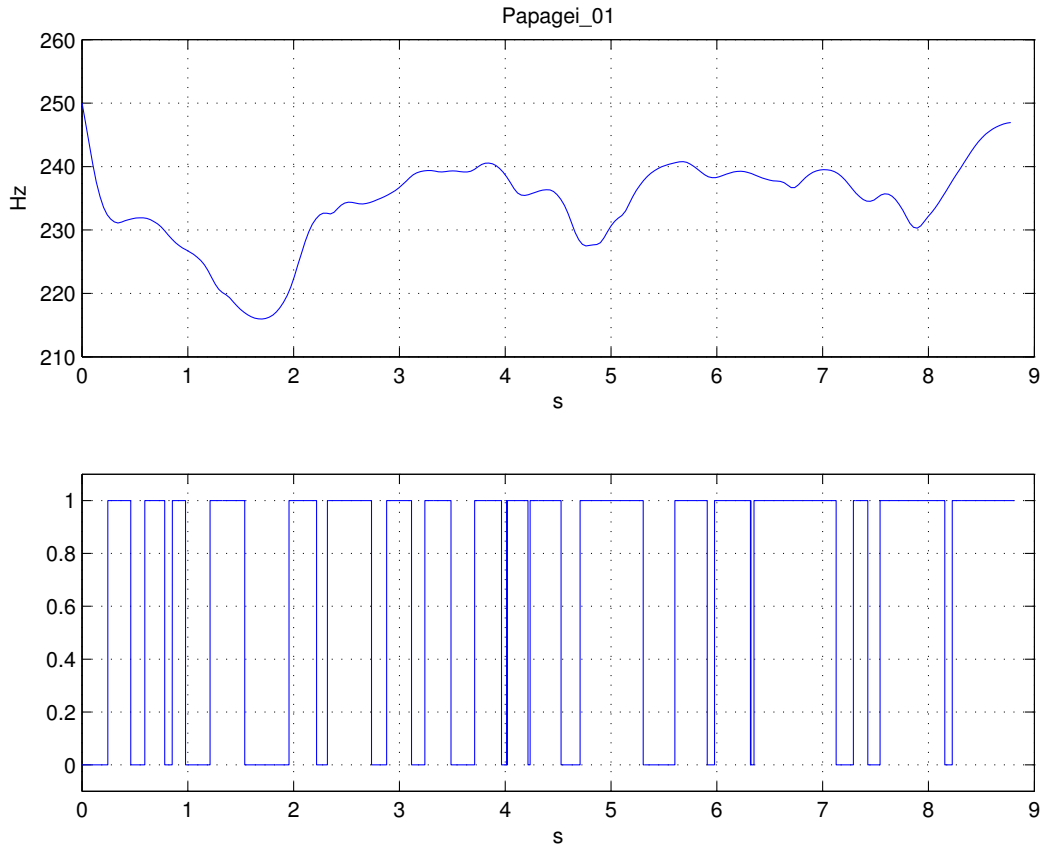


Figur 3.3: *Ausschnitt eines gekrächzt gesprochenen [e] (oben) sowie die entsprechenden Fehlersignale ohne (mitte) und mit Präemphase (unten). Wir verwendeten ein Hanning-Fenster und eine Prädiktorordnung von 60. Insbesondere im Fehlersignal ohne Präemphase ist die aperiodische Anregung gut sichtbar.*

Nichtsdestotrotz ist die gehörmässige Beurteilung des Dekonstruktionsversuches (Hörbeispiel 3.3) nicht trivial, da zu dem Charakter Papagei bzw. Rabe noch diverse andere Phänomene wie Formantverschiebungen oder die spezielle Prosodie hinzukommen. Diese verdecken oder überlagern zumindest das Krächzen in der subjektiven Wahrnehmung zu einem guten Teil. Zudem erzeugt die LPC-Synthese nicht unbedeutende Artefakte, die selbst durchaus eine gewisse Ähnlichkeit zum Krächzen haben. Trotzdem können wir den Dekonstruktionsversuch an dieser Stelle als gelungen bezeichnen.

In Figur 3.6 ist das Verschwinden der Aperiodizität deutlich sichtbar und bestätigt die bisherigen Erkenntnisse.

Die für den Dekonstruktionsversuch benötigte Funktionalität stellt die Matlab-Funktion `LPC_ANALYSESYNTHESE_FO03` zur Verfügung; die Implementation wird im Anhang A.7 genauer beschrieben.



Figur 3.4: *Für den Dekonstruktionsversuch verwendeter Grundfrequenzverlauf eines gekrächzt gesprochenen Satzes (oben) und der dazugehörige Stimmhaft-/Stimmlos-Entscheid (unten).*

3.2.7 LPC-Analyse/Synthese als Konstruktionsversuch

Von besonderem Interesse war an dieser Stelle natürlich auch ein Konstruktionsversuch; als erstes einfaches Vorgehen verwendeten wir einen Algorithmus, der zwischen den anregenden Impulsen der LPC-Synthese Nebenimpulse setzt. Zu Beginn wählten wir die Position der zusätzlichen Pulse exakt zwischen zwei Hauptimpulsen und die Grösse gemäss einem fixen Bruchteil des Hauptimpulses. Mittels eines zufälligen Verfahrens bestimmten wir, ob an einer gegebenen Position ein Impuls gesetzt wird oder nicht.

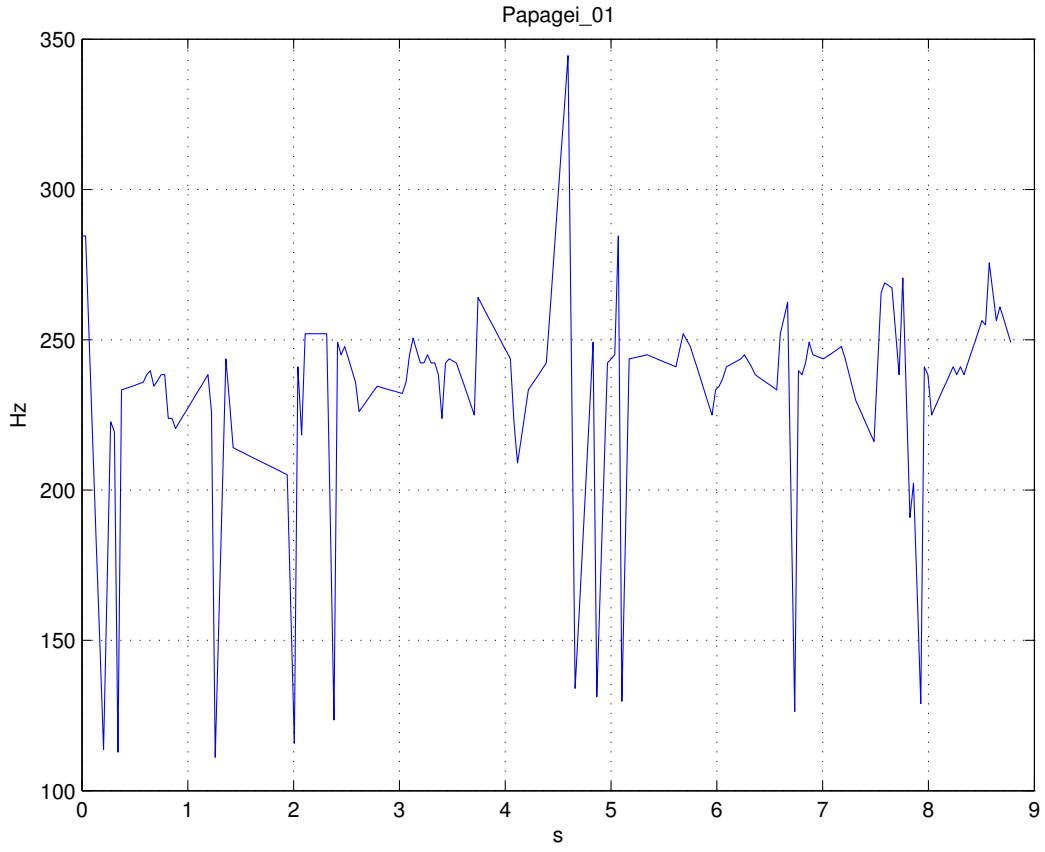
Dieses Vorgehen brachte eine leichte Verhärtung des Klanges mit sich, von einem eigentlichen Krächzen konnte jedoch noch nicht die Rede sein. Das zufällige Bestimmen der Impulsgrösse brachte keine Verbesserung dieses Resultates.

Ein erfolgreicheres Resultat erzielten wir, indem wir zusätzlich die Position des Zwischenimpulses zufällig variierten. Die damit berechneten Synthesesignale zeigen eindeutig eine gewisse Krächz-Charakteristik. Insgesamt erhielten wir jedoch eher den Eindruck technischer Artefakte als eines natürlichen Krächzens. Die Implementation ist in der Matlab-Funktion `LPC_SYNTHESIS_MOD` zu finden.

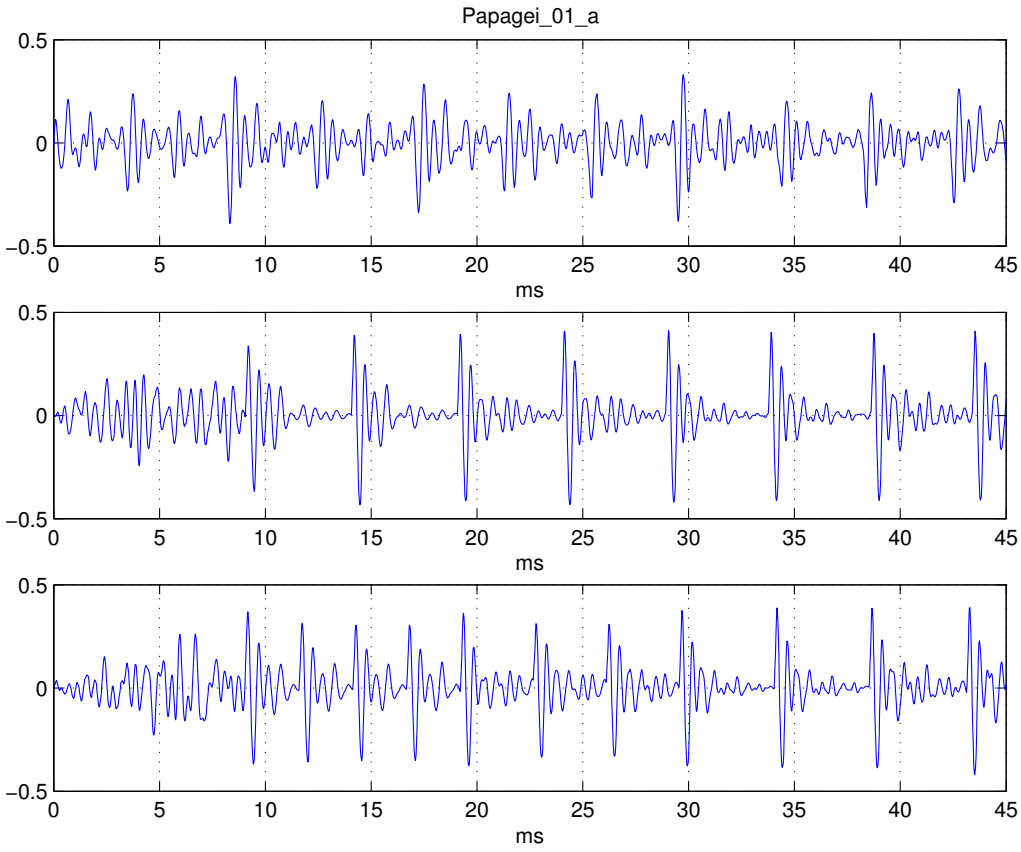
Für einen genaueren Konstruktionsversuch war es offenbar nötig, die im Orginalsignal vorhandene Aperiodizität nachzubilden; Basis dafür war ein von der Sprachverarbeitungsgruppe zur Verfügung gestellter Algorithmus, der dieses Verhalten mittels eines stark schwankenden Grundfrequenzverlaufs modelliert. Die Grundfrequenzdetektion basiert dabei auf einem heuristischen Verfahren im Zeitbereich (siehe Figur 3.5). Verwendet man diesen Grundfrequenzverlauf zusammen mit einer passenden Fenstergrösse (wir wählten 1200 Samples bei 44.1 kHz Samplefrequenz) in der LPC-Synthese, ergibt sich eine erstaunlich realistisch Nachbildung des Krächzens (Hörbeispiel 3.4).

Figur 3.6 zeigt den Vergleich von Orginalsignal, Dekonstruktions- und Konstruktionsversuch im Zeitbereich. Die Pulsposition und insbesondere die Impulsamplitude entspricht im Konstruktionsversuch nicht exakt dem Orginalsignal, was aber gehörmässig wenig Einfluss zu haben scheint.

Für diesen Versuch verwendeten wir wiederum die Matlab-Funktion `LPC_ANALYSESYNTHESE_FO03`, die im Anhang detailliert beschrieben ist.



Figur 3.5: *Mit der heuristischen Grundfrequenzdetektion im Zeitbereich erhaltener Grundfrequenzverlauf.*



Figur 3.6: *Orginalsignal eines gekrächzten Lautes [a] (oben), Signal des Dekonstruktionsversuches (mitte) und Signal des Konstruktionsversuches (unten). Sichtbar ist im Orginalsignal die aperiodische Anregung, die im Dekonstruktionsversuch nicht modelliert und durch eine systematische Impulsfolge ersetzt wird. Im Konstruktionsversuch erkennen wir die stark schwankenden Impulsabstände, was auf die korrekte Modellierung der Aperiodizität hinweist.*

3.3 Resultate

Zusammenfassend können wir die aperiodische Anregung als Hauptmerkmal für eine krächzende Stimme bezeichnen. Für die entsprechenden Tiercharaktere Papagei bzw. Rabe scheinen noch weitere Merkmale wie die Anhebung bei 2 kHz sowie der konstante Grundfrequenzverlauf bzw. allgemein die Prosodie dazuzukommen.

3.4 Weitere Versuche

Im folgenden zeigen wir wiederum mögliche weiterführende Versuche auf.

3.4.1 Impulsstärke korrigieren

Wie in Figur 3.6 ersichtlich, stimmt die Impulsstärke gewisser Nebenpulse nicht; es ist zu vermuten, dass der gewünschte Effekt dadurch leicht abgeschwächt wird. Eine Korrektur ist u.U. durch eine leichte Anpassung der LPC-Analyse (z.Bsp. mit einer veränderten Fensterlänge) möglich. Ansonsten kann die Amplitude einfach in der Funktion zur aperiodischen Grundfrequenzdetektion bestimmt und in die LPC-Synthese übernommen werden.

3.4.2 Systematik der Anregung analysieren

Für eine Resynthetisierung von normal gesprochenen Sprachsignalen mit Krächzstimme wird ein Modell der aperiodischen Anregung benötigt; in Figur 3.5 ist ersichtlich, dass der Algorithmus für die Grundfrequenzdetektion für viele Passagen einen Wert von ca. 240 Hz liefert, was auch ungefähr der gehörmässigen Wahrnehmung entspricht. Die unstetigen Anteile bewegen sich typischerweise bei der halben Frequenz, also 120 Hz. Es ist durchaus möglich, dass aufgrund dieser Beobachtung eine einfache Systematik in der Anregung gefunden werden kann.

3.4.3 Weitere Merkmale berücksichtigen

Für eine Resynthetisierung mit Papageien- bzw. Rabenstimme ist ebenfalls die Modellierung des veränderten Spektrums – also die Anhebung bei 2 kHz – und der Prosodie nötig. Betreffend der Prosodie (insbesondere des Grundfrequenzverlaufs) sind jedoch weitergehende Unstersuchungen nötig.

A Matlab-Funktionen

Im folgenden geben wir für jede auf der CD vorhandene Matlab-Funktion einen kurzen Funktions- und Verwendungsbeschrieb.

A.1 Funktionen im Verzeichnis `basic_functions/`

`Y=ADD_ARTEFACTS01(sig,smooth,t_strength,t_speed)` modifiziert ein Signal durch Tiefpassfilterung und Hinzufügen von Rauschen. Falls der Parameter `smooth` grösser als 0.001 ist, wird das Signal phasenlinear tiefpassgefiltert. Ist er kleiner, wird das Signal nicht gefiltert. `t_strength` steuert die Amplitude des zugemischten Rauschens und `t_speed` die Ansatzfrequenz des Tiefpassfilters, der das Rauschen vor der Addition filtert.

Diese Funktion benötigten wir insbesondere, um Grundfrequenzverläufe zu glätten.

`[pulse,t]=APERIODIC_BASEFREQUENCY(sig,Fs)` dient der Berechnung der Grundfrequenz bei aperiodisch angeregten Sprachsignalen `sig` mit Samplefrequenz `Fs`. Verwendung fand diese Funktion beim Konstruktionsversuch der Krächzstimme.

`Y=CONVERT_LIST01(list,length)` konvertiert eine aus Positions-/Wertepaaren bestehende Variable `list` durch lineare Interpolation in eine gleichabständige Wertefolge der Länge `length`. Wird benötigt, um bei der LPC-Analyse eigene Grundfrequenz-, Dauer- und Amplitudenverläufe vorzugeben.

`Y=CUT_TO_ZERO01(sig)` beschneidet das Signal `sig` am Anfang und am Ende um alle Werte, die gleiches Signum wie das erste bzw. letzte Sample haben. Diese Funktion haben wir beim Zuschneiden eines Signalausschnittes auf eine exakte Periodenzahl sowie im Zusammenhang mit der Funktion `GET_CUTOUT03` verwendet.

`[X,B,A]=FILTERBOX03(sig,frequencies,amplifications [,Fs])` filtert das Eingangssignal `sig` mittels eines Filters, das Anhand der vorgegebenen Frequenzpositionen `frequencies` und zugehörigen Verstärkungswerten `amplifications` designed wird. Als Rückgabewert erhalten wir das gefiltere Signal `X` sowie die Koeffizienten `B` und `A` des Filters.

Diesen Programmcode entwickelten wir für den Konstruktionsversuch der gepressten Stimme.

`[Y,Fo]=FREQ_DIFF01(file1,file2,label1 [,label2 [,label3 [,label4]]])` berechnet die Differenz des gemittelten geglätteten Spektrums aller in `file1` und `file2` gelisteten WAV-Files; die Mittelung wird nur über die Laute `label1` bis `label4` vorgenommen. Diese Funktion benötigten wir ebenfalls für den Konstruktionsversuch der gepressten Stimmme.

`[Y,Fs]=GET_CUTOUT03(file,label)` gibt alle Instanzen des Lautes `label` im WAV-File

`file` zurück; wird in diversen Funktionen verwendet.

`Y=GET_VOICING01(sig,Fs)` verwendet eine einfache Schwellenwertdetektion für das Sprachsignal `sig`. Es wird der RMS-Pegel verwendet, um auch bei verstellten Stimmen einen robusten Stimmhaft-/Stimmlos-Entscheid zu erhalten. Diesen verwendeten wir für die LPC-Analyse der Krächzstimme.

`Y=LPC_ANALYSESYNTHESE_F003(file)` erzeugt für ein WAV-File `file` den Grundfrequenzverlauf sowie den Stimmhaft-/Stimmlos-Entscheid. Mit diesen Parametern wird anschliessend eine LPC-Analyse/Synthese aufgerufen. Diese Funktionalität benötigten wir für den Konstruktions- und Dekonstruktionsversuch der Krächzstimme.

`syn=LPC_VARS04(sig [,f_list [,v_list [,d_list [,a_list]]]])` ist eine LPC-Analyse/Synthese für das Signal `sig`, wobei für den Grundfrequenzverlauf, den Stimmhaft-/Stimmlos-Entscheid, die Dauer sowie die Amplitude eigene Kontrollsignale vorgegeben werden können. Die Kontrollsignale werden in Form von Zeit-/Wertepaaren übergeben. Zusätzlich ermöglicht die Funktion eine Vokaltraktänderung durch Resamplen des Signals.

Diese Variante der LPC-Analyse/Synthese benutzten wir für unterschiedliche Konstruktions- und Dekonstruktionsversuche.

`Y=NORMALIZE_LOUDNESS01(sig,Fs)` normalisiert die Lautstärke des Eingangssignales `sig` anhand des maximalen RMS-Pegels des hochpassgefilterten Signals. Die Hochpassfilterung dient der Nachbildung des Frequenzganges des menschlichen Gehörs. Diese Normalisierung verwendeten wir insbesondere bei den statistischen Analysen.

`PLOT_DUMMIES01` erzeugt neun Plots von gekrächzt und unverstellt gesprochenen Sprachsignalen, die den Zeit- und Frequenzbereich des jeweiligen Signalausschnittes zeigen. Diese Darstellungen dienten uns der Analyse der Krächzstimme.

`X_win_error=PLOT_FEHLERSIGNAL01(file,label)` plottet die Zeitbereichsdarstellung des Lautes `label` aus dem WAV-File `file` sowie das Fehlersignal des entsprechenden Ausschnittes mit und ohne Präemphase. Zusätzlich wird das Spektrum des Fehlersignals ohne Präemphase angezeigt. Verwendung fanden diese Plots in der Analyse der aperiodischen Anregung bei Krächzstimmen.

`PLOT_FORMANTS04(file,label [,Cepwin [,opt]])` erstellt für das cepstral geglättete Spektrum des Lautes `label` des WAV-Files `file` eine Colormap- und/oder Wasserfalldiagrammdarstellung. Sie dienten uns ersten Untersuchungen der gepressten Stimme.

`PLOT_SIGNAL_AND_SPEC(sig,Fs,sec)` wird von `PLOT_DUMMIES01` verwendet, um ein Signal `sig` im Zeit- und Frequenzbereich darzustellen. Der Parameter `sec` steuert dabei die Einheit der Skala der x-Achse (Samples oder Millisekunden).

`Z=RESAMPLER01(sig,f1,f2)` resampelt ein Eingabesignal `sig` mit Samplefrequenz `f1` auf eine Samplefrequenz `f2`. Als Interpolationsverfahren verwenden wir ein Chebyshev-Tiefpassfilter hoher Ordnung, eingesetzt in einer Filterroutine für linearphasige Filterung.

Diese Funktionalität benötigten wir für experimentelle Vokaltraktänderungen.

`f_opt=SEARCH_GOOD_FREQ01(f1,f2,tolerance)` berechnet im durch `tolerance` als Prozentwert angegebenen Umfeld von `f2` einen neuen Wert `f_opt`, der bezüglich `f1` minimales kleinstes gemeinsames Vielfaches hat. Diese Optimierung verwendeten wir vor dem Aufruf von `RESAMPLER01`, um die Rechenzeit in vernünftigen Grenzen zu halten.

`Y=STAT_ANALYSE04(file,label [,label2 [,label3 [,label4]]])` berechnet eine statistische Spektralanalyse über mehrere im Textfile `file` gelistete Sprachsignale und bis zu vier Lauten `label` bis `label4`. Die Spektren werden dabei cepstral geglättet.

`[m,fo]=WEIGHTED_FREQ01(F_end,file,label1 [,label2 [,label3 [,label4]]])` ist eine leicht modifizierte Version von `STAT_ANALYSE04`, die wir für den Konstruktionsversuch der gepressten Stimme benötigten.

A.2 Funktionen im Verzeichnis `basic_functions/utilities/`

Im Verzeichnis `basic_functions/utilities/` haben wir alle Funktionen abgelegt, die grundlegende Signalverarbeitungsfunktionen zur Verfügung stellen.

`[G,mx,P,A]=LPC_analysis(s,p,Pmin,Pmax,vthr)` ist die von der Sprachverarbeitungsgruppe zur Verfügung gestellte Funktion zur LPC-Analyse. Wir verweisen für nähere Angaben auf [1].

`[so,do,zo]=LPC_SYNTHESIS(G,P,di,A,ns,zi)` ist die von der Sprachverarbeitungsgruppe zur Verfügung gestellte Funktion zur LPC-Synthese. Wir verweisen dazu auf [1].

`[so,do,zo]=LPC_SYNTHESIS_MOD(G,P,di,A,ns,zi,mod)` ist eine modifizierte Version der Funktion `LPC_SYNTHESIS`; sie diente einem ersten Konstruktionsversuch der gekrächzten Stimme. Zur aperiodischen Anregung setzt sie zufällig Zwischenimpulse zu den Hauptimpulsen der gegebenen Grundfrequenz. Wird die Variable `mod` auf 0 gesetzt, werden keine Zwischenimpulse gesetzt.

`[A,G,r,a]=AUTOLPC(x,p)` berechnet die LPC-Koeffizienten mit einer Autokorrelationsmethode und wurde uns von der Sprachverarbeitungsgruppe zur Verfügung gestellt. Für nähere Angaben verweisen wir auf [1].

`Y=GET_SINUS(f,n)` gibt eine Sinuswelle der Frequenz `f` (bzgl. der Samplefrequenz 44.1 kHz)

und der Länge `n` Samples zurück.

`Y=MAKE_MONO(sig)` konvertiert Stereo-Audiosignale in Monosignale. Dabei wird der Kanal mit dem höheren RMS-Pegel ausgewählt.

`PLAY_SIG(sig,Fs)` dient uns zum Abspielen von Audiosignalen. Vor der Ausgabe wird das Signal so normalisiert, dass der maximale Pegel 6 dB unterhalb der Clippgrenze liegt.

`Xfft=PLOT_SPEC(sig [,Fs [,opt [,color [,window]]]])` stellt Signale im Frequenzbereich dar. Es stehen folgende Optionen zur Verfügung: lineare oder logarithmische Darstellung der Frequenzachse, Farbe des Plots und zuschaltbare Fensterfunktion.

`Y=RMS_DETECTOR(sig,TAV)` berechnet den RMS-Wertverlauf des Signales `sig` mit der Zeitkonstante TAV mittels eines Leaky Integrators.

A.3 Funktionen im Verzeichnis `basic_functions/old/`

In diesem Verzeichnis befinden sich nicht mehr benötigte Funktionen. Sie sind nur teilweise ausführlich dokumentiert.

A.4 Skripts im Verzeichnis `bericht/bilder/`

Die für die Generierung der in diesem Bericht gezeigten Darstellungen verwendeten Matlab-Skripts befinden sich im Verzeichnis `bericht/bilder/`. Diese Skripts verwenden zu einem grossen Teil verschiedene Funktionen aus den Verzeichnissen `basic_functions/` und `basic_functions/utilities/`; um die Skripts korrekt ausführen zu können, sollten aus diesem Grund die entsprechenden Pfade im Matlab Preferences-File angegeben werden.

A.5 Skripts im Verzeichnis `bericht/hoerbeispiele/`

Im Verzeichnis `bericht/hoerbeispiele/` sind die Skripts abgelegt, die die Hörbeispiele generieren und abspielen. Sie benötigen ebenfals die unter A.4 angegebenen Pfade im Matlab Preferences-File.

A.6 Diskussion der Funktion `STAT_ANALYSE06`

Im folgenden werden wir die Implementation und Benutzung der Funktion `Y=STAT_ANALYSE04(file,label [,label2 [,label3 [,label4]]])` diskutieren. Verwendung fand diese Routine bei der Berechnung der statistischen Spektralanalyse, die wir für die Untersuchung der gepressten Stimme benötigten.

Die Variable `file` enthält als String den Namen bzw. Pfad eines Textfiles mit der Endung `*.stat` (die im Namensstring nicht angegeben werden sollte). Diese Textdatei enthält ein Listing von Pfaden zu WAV-Dateien, die für die Statistik verwendet werden.

Es wird angenommen, dass die Labeldateien im gleichen Verzeichnis wie die zugehörigen Audiodateien gespeichert sind und denselben Namen, aber die Endung `*.lab` besitzen. Wir geben hier ein Beispiel des Inhaltes einer solchen `*.stat`-Datei:

```
../dae_chaschper/Riese_01
../dae_chaschper/Riese_02
../dae_chaschper/Riese_03
../saetze/Riese_01
../saetze/Riese_02
../saetze/Riese_03
../saetze/Riese_04
```

Die im Prozeduraufruf folgenden Variablen enthalten Strings, nach denen die entsprechenden Labels in den Labelfiles gesucht werden. Es können maximal vier verschiedene Labelstrings angegeben werden.

Die Prozedur iteriert in einer `while`-Schleife über alle im `*.stat`-File gelisteten Audiodateien. Mit einem Aufruf der Funktion `GET_CUTOUT03` erhalten wir in einer Matrix alle im aktuellen Audiofile auftretenden Instanzen aller angegebenen Labels.

Jede Instanz wird zuerst bezüglich Lautheit normalisiert, um eine aussagekräftige Mittelwertsbildung zu ermöglichen. Anschliessend berechnen wir das cepstral geglättete Spektrogramm jeder Labelinstanz; für jedes Spektrogramm bilden wir im folgenden den zeitlichen Mittelwert. Um dabei Randeffekte auszublenden, werden die mittleren Spektren unter Verwendung eines Hamming-Fensters stärker gewichtet. Wie Figur A.1 zeigt, sind die Unterschiede zu einer ungewichteten Mittelwertbildung im Endresultat jedoch marginal. Dies zeigt die Robustheit der Methode.

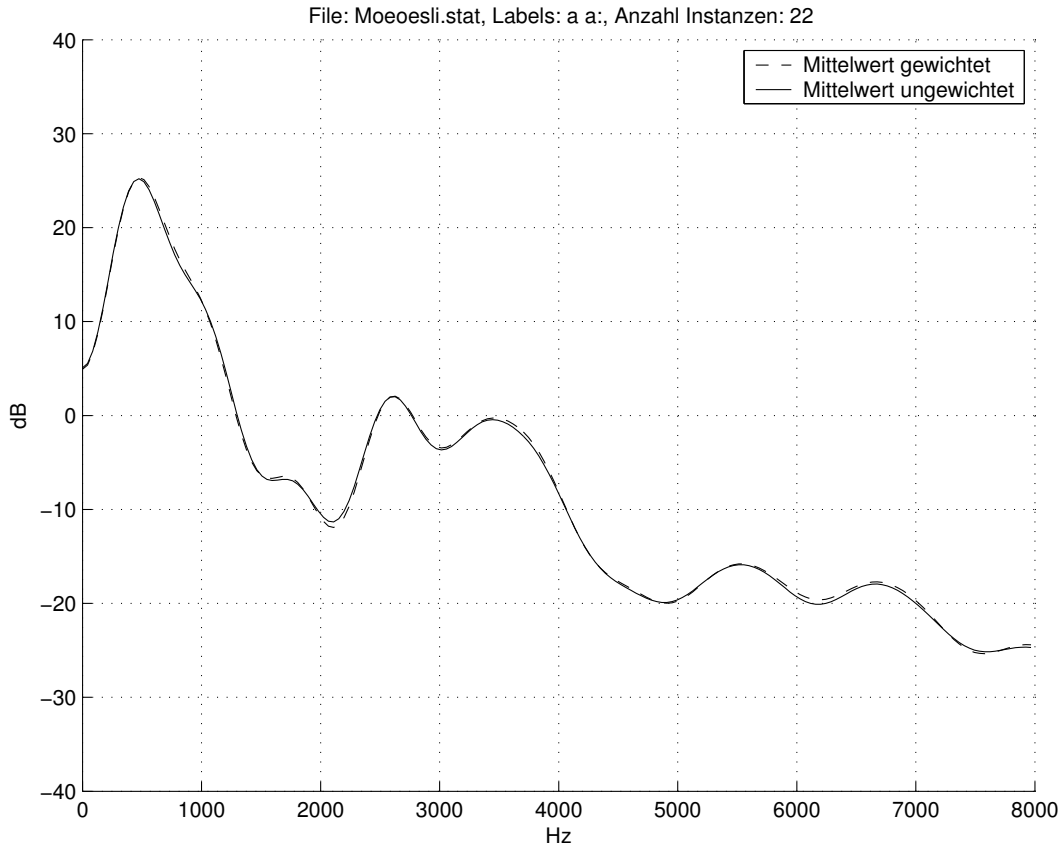
Das so erhaltene Spektrum speichern wir als Zeileneintrag in eine Matrix; für jede Labelinstanz erhalten wir damit einen Eintrag in der Matrix, von der wir am Ende der Berechnung den Mittelwert und die Standardabweichung bestimmen und darstellen.

A.7 Diskussion der Funktion `LPC_ANALYSESYNTHESE_FO03`

Die Funktion `LPC_ANALYSESYNTHESE_FO03(file)` benötigten wir für den Konstruktions- und Dekonstruktionsversuch der gekrächzten Sprechstimme. Die Variable `file` enthält den Filenamen einer Audiodatei, für die eine LPC-Analyse/Synthese ausgeführt wird.

Der Grundfrequenzverlauf stammt dabei entweder von der Funktion `GETBASEFREQ` oder `APERIODIC_BASEFREQUENCY`. Erstere verwenden wir für den Dekontruktions-, zweitere für den Konstruktionsversuch. Beide wurden uns von der Sprachverarbeitungsgruppe zur Verfügung gestellt. Die Umstellung von Konstruktions- auf Dekonstruktionsversuch ist im Matlab-Sourcecode beschrieben.

Für den Stimmhaft/Stimmlos-Entscheid verwenden wir die Funktion `GET_VOICING01`. Anschliessend werden die beiden Kontrollsignale in ein spezifisches Listenformat konvertiert und der Routine `LPC_VARS04` übergeben.



Figur A.1: *Gemitteltes Spektrum, mit und ohne zeitliche Gewichtung durch ein Hamming-Fenster. Es ist klar ersichtlich, dass die Unterschiede sehr gering ausfallen.*

Da das bestimmen der Grundfrequenz mit der Funktion `GETBASEFREQ` rechen- und damit zeitaufwendig ist, lassen sich die entsprechenden Variablen speichern und für spätere Berechnungen wieder laden.

B Beschreibung der beiliegenden CD

Die beiliegende CD ist in folgende drei Hauptverzeichnisse aufgeteilt:

```
basic_functions/
bericht/
chaschperli/
```

Im Verzeichnis `basic_functions/` sind alle von uns programmierten Matlab-Funktionen gespeichert. Sie sind übersichtsweise im Anhang A beschrieben.

Im zweiten Verzeichnis findet sich dieser Bericht als PDF-Datei. Zusätzlich sind die Hörbeispiele (Matlabskripte und WAV-Dateien) und die Skripts zu den Figuren in den Unterverzeichnissen `hoerbeispiele/` bzw. `bilder/` zu finden.

Die Orginalaudiodaten der Kasperli-CDs sind im Verzeichnis `chaschperli/` abgelegt. Sie liegen als 16 bit stereo, 44.1 kHz WAV-Dateien vor.

C Aufgabenstellung

Institut für Technische Informatik und Kommunikationsnetze
ETH Eidgenössische Technische Hochschule Zürich, Swiss Federal Institute of Technology Zurich, Ecole polytechnique fédérale de Zurich, Politecnico federale di Zurigo
Gruppe für Sprachverarbeitung

Wintersemester 2004/2005

SEMESTERARBEIT

für

Samuel Groner und Andreas Scherrer

Betreuer: H. Romsdorfer ETZ D97.5
Stellvertreter: Dr. B. Pfister ETZ D97.6

Ausgabe: 18. Oktober 2004
Abgabe: 4. Februar 2005

Sprache mit Tierstimmencharakteristik

Einleitung

Bei Sprachsynthesesystemen basierend auf dem Verkettungsansatz (d.h. durch Verkettung von aufgenommenen Sprachgrundelementen) hat die resultierende Stimme mehr oder weniger den Charakter des originalen Sprechers der Sprachgrundelemente. Es existieren mittlerweile einige erfolgversprechende Versuche mittels Übertragung der Prosodie und der segmentalen Charakteristik der resultierenden Stimme den Stimmcharakter eines anderen Sprechers aufzuprägen.

In manchen Anwendungsgebieten, wie etwa in Computerspielen oder beim Vorlesen von Märchen, wäre es jedoch wünschenswert, der zu erzeugenden Stimme die Stimmeigenschaften eines bestimmten, fiktiven Charakters (z.B. eines Hundes, eines Raben, eines Riesen oder etwa eines Drachen) verleihen zu können. Natürlich kann man dabei nicht von einer Stimmkonversion im anfangs erwähnten Sinn sprechen. Eher kommt es darauf an, bestimmte Effekte (wie etwa Krächzen, Zischen, etc.), welche mit dem gewünschten Charakter assoziiert werden, der zu erzeugenden Stimme aufprägen zu können.

Problemstellung

Im Rahmen dieser Arbeit sollen für bestimmte, ausgewählte fiktive Charaktere die wesentlichen Stimmeigenschaften (wie z.B. Krächzen, Zischen, usw.) identifiziert und ana-

lysiert werden. Dazu ist es notwendig, geeignete Beispielaufnahmen von Sprechern (wie z.B. Märchenerzählern) zu sammeln. Optimal wären in diesem Fall Aufnahmen von unterschiedlichen Charakteren, gesprochen vom gleichen Sprecher.

In einem weiteren Schritt sollen die lautlichen Eigenschaften der Beispielaufnahmen gleicher Charaktere untereinander verglichen werden und Gemeinsamkeiten bestimmt werden. Eventuell könnte man auch im Fall von Tiercharaktere Ausschnitte von Aufnahmen mit imitierten Tierstimmen zusätzlich mit ähnlichen, echten Tieraufnahmen vergleichen. Durch Vergleich der Unterschiede zwischen verschiedenen Charakteren sollen dann die einen bestimmten Charakter spezifizierenden lautlichen Eigenschaften herausgefunden werden.

Nach Bestimmung der speziellen lautlichen Eigenschaften könnten, je nach Machbarkeit und Aufwand, für einige ausgewählte Charaktere Transformationsverfahren entwickelt werden, um Sprachaufnahmen von einer menschlichen Stimme mit dem gewählten Charakter versehen zu können.

Aufgaben

1. Einarbeitung in die Literatur und in die Arbeitsumgebung (für eine Einführung in die Sprachverarbeitung siehe [1]).
2. Sammlung von geeigneten Sprachbeispielen.
3. Identifikation und Analyse von speziellen Stimmeigenschaften der unterschiedlichen Charaktere. Im Speziellen Herausarbeitung der Gemeinsamkeiten von Aufnahmen gleicher Charaktere und der Unterschiede zwischen Aufnahmen verschiedener Charaktere.
4. Der Nachweis einer bestimmten Stimmeigenschaft kann mit Hilfe von Konstruktions- und Destruktionsexperimenten an den gesammelten Sprachbeispielen erbracht werden.
5. Die ausgeführten Arbeiten und die erhaltenen Resultate sind in einem Bericht (siehe Richtlinien unter "www.tik.ee.ethz.ch/~spr/SADA/") zu dokumentieren, der in zwei Exemplaren abzugeben ist, wovon eines Eigentum des Instituts bleibt.

Literaturverzeichnis

[1] B. Pfister und H.-P. Hutter. *Sprachverarbeitung I.* Vorlesungsskript für das Wintersemester 2003/2004, Departement ITET, ETH Zürich, 2003.

Zürich, den 18. Oktober 2004 Prof. Dr. L. Thiele

Literatur

[1] Beat Pfister und Hans-Peter Hutter: *Sprachverarbeitung I*, Vorlesung und Vorlesungsskript für das Wintersemester 2003/2004, Departement ITET, ETH Zürich, 2003.