Ronald Schulz

Bandbreitenerweiterung von schmalbandigen Sprachsignalen

Projektarbeit

vorgelegt an der
Technischen Universität Graz und der Universität für Musik und Darstellende Kunst, Graz

Betreuer: DI Markus Noisternig

Institut für Elektronische Musik und Akustik

Juni 2007

Kurzfassung

Sprachsignale beinhalten Frequenzen von ungefähr 100 Hz bis 8 kHz (breitbandige Sprache). In Telekommunikationsanlagen wird Sprache in einem Frequenzbereich von 300 Hz bis 3.4 kHz (schmalbandige Sprache) übertragen. Die demzufolge schlechtere Signalqualität führt zu einer Beeinträchtigung der Sprachverständlichkeit. Das macht sich vor allem bei Frikativen, wie /f/ oder /s/ bemerkbar, bei denen ein großer Anteil der Information im Frequenzband von 3.4 kHz – 8 kHz liegt. Ziel dieses Projekts ist es, ein Programm in MATLAB zu implementieren, das aus einem schmalbandigen Sprachsignal das obere Frequenzband schätzt. Dadurch soll die Sprache natürlicher klingen und die Verständlichkeit erhöht werden.

Abstract

Speech signals contain frequencies between 100 Hz and 8 kHz (broadband speech). In telecommunication systems speech is transmitted in a frequency range between 300 Hz and 3.4 kHz (narrowband speech). This leads to a degradation of the intelligibility of the speech, since an important part of the frequency content of fricatives like /f/ or /s/ are in the upper band (3.4 kHz - 8 kHz). This project aims at the implementation of an algorithm in MATLAB to estimate the upper band out of the narrow band. Consequently, the speech sounds more natural and the intelligibility is being increased.

Abkürzungen

BWE - Bandbreitenerweiterung *(bandwidth extension)*
CEPD - Cepstral Distance
DIS - Itakura Saito Distance
LPC - Linear Predictive Coding
LSD - Log Spectral Distortion

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die menschliche Sprache beinhaltet Frequenzen im Bereich von ungefähr 100 Hz bis 8 kHz. Beim Telefonieren mit herkömmlichen analogen Telefonanlagen wird das Sprachsignal nur im Frequenzbereich von 300 Hz bis 3400 Hz vollständig übertragen [siehe ITU-T Norm G.120]. Da jedoch Frikative, wie /s/, /sch/, /f/ oder Plosivlaute, wie /p/, wichtige Informationen im Frequenzband von 3.4 kHz bis 8 kHz enthalten, kommt es durch die Signalübertragung zu einer Verschlechterung der Verständlichkeit.
Untersuchungen haben gezeigt, dass es einen Zusammenhang zwischen dem unteren und oberen Frequenzband gibt [1]. Deshalb ist es möglich, rein aus dem unteren Frequenzband von 300 Hz bis 3.4 kHz, das obere Frequenzband zu schätzen, ohne zusätzliche Informationen davon zu benötigen. Das bringt den Vorteil einer Verbesserung der Sprachverständlichkeit bei verhältnismässig geringem technischem Aufwand. Die Modifikationen müssen nur empfangsseitig durchgeführt werden.
In dieser Projektarbeit wurde mit MATLAB eine mögliche Art der Implementierung der Bandbreitenerweiterung (*bandwidth extension*- BWE) realisiert. Basierend auf dem Quelle- Filter Modell der Spracherzeugung wird diese BWE mit der Methode des *Linear Mapping* durchgeführt.
Zu Beginn dieser Arbeit werden kurz einige Grundlagen der Sprachsignalverarbeitung, wie zum Beispiel das Quelle- Filter Modell, erläutert. Im darauffolgenden Kapitel geht es allgemein um die theoretischen Grundlagen der BWE. Im 3. Kapitel wird das implementierte System vorgestellt und die einzelnen Komponenten beschrieben. Für die Auswertung der Ergebnisse sind dann noch einige Fehlermaße notwendig, um objektive Vergleiche zwischen Original und geschätztem Signal ziehen zu können. Am Schluss werden die Ergebnisse vorgestellt und es wird einen Ausblick auf mögliche Verbesserungen des Systems geben.

Kapitel 2

Grundlagen

2.1 Lineares Modell der Spracherzeugung

Wenn wir sprechen, strömt Luft durch die Stimmritze (Glottis) und die Stimmbänder fangen an sich zu bewegen. Je nachdem, ob zum Beispiel ein Vokal oder ein Konsonant erzeugt werden soll, schwingen die Stimmbänder unterschiedlich. Dadurch erzeugen wir ein sogenanntes Anregungssignal, welches nun noch im Vokaltrakt (Rachen- und Mundraum) spektral geformt wird. Das geformte Signal wird in weiterer Folge von den Lippen und den Nasenlöchern abgestrahlt. Das lineare Modell der Spracherzeugung versucht die menschliche Spracherzeugung nachzubilden. In [2, S. 8 ff.] ist eine genaue Beschreibung dieses Modells zu finden.
Grundlage dieser Projektarbeit ist das vereinfachte lineare Quelle- Filter Modell der Spracherzeugung. Es beruht auf einen rein rekursiven Filter, der also nur Polstellen besitzt. Man verwendet deshalb auch den Begriff Autoregressiver (AR) Prozeß. Die Modellierung des Nasal- Traktes kann durch ein Pol- Nullstellen Filter realisiert werden. Es hat sich aber herausgestellt, dass der Nasal- Trakt, sowie der Glottis- und Lippen- Filter aus dem ursprünglichen Modell für Anwendungen in der Sprachsynthese vernachlässigt werden können, ohne dass es zu nennenswerten Qualitätseinbußen kommt [2, S. 167]. Das lässt sich wie folgt erklären:
Bei der Übertragungsfunktion $H(z)$ eines stabilen und kausalen Pol- Nullstellen Filters liegen sämtliche Polstellen innerhalb des Einheitskreises. Die Nullstellen können aber auch ausserhalb des Einheitskreises liegen. In diesem Fall lässt sich $H(z)$ in einen minimalphasigen Anteil $H_{min}(z)$ und eine Allpass- Übertragungsfunktion $H_{Ap}(z)$ laut

$$H(z) = H_{min}(z)H_{Ap}(z) \tag{2.1}$$

aufspalten. Für die Sprachsynthese ist es ausreichend, den minimalphasigen Anteil zu realisieren, da unser Ohr gegenüber Phasenänderungen, hervorgerufen vom Allpass, weitgehend unempfindlich ist. Da nun sämtliche Pole und Nullstellen des

minimalphasigen Filters innerhalb des Einheitskreises liegen, gibt es auch einen stabilen inversen Filter mit

$$H_{min}^{-1}(z) = \frac{1}{H_{min}(z)}. \tag{2.2}$$

Dadurch lässt sich die durch den Vokaltrakt hervorgerufene Filterung wieder rückgängig machen und somit das Anregungssignal des Vokaltraktes zurückgewinnen. Darüber hinaus kann jedes minimalphasige Pol- Nullstellen Filter mit einem Allpol-Filter m- ten Grades angenähert werden. Damit erklärt sich die Verwendung eines Allpol- Filters bei der Sprachsynthese [2, S. 167 ff.]. Das vereinfachte lineare Quelle- Filter Modell ist in 2.1 abgebildet.

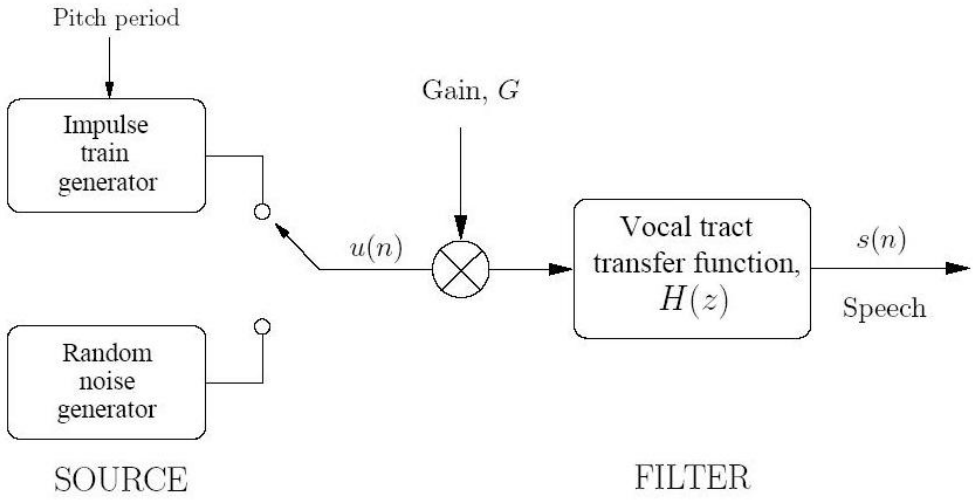

Abbildung 2.1: Quelle- Filter Modell [3]

Entsprechend der Physiologie des menschlichen Sprechapparates kann das Modell in zwei primäre funktionelle Verarbeitungsstufen getrennt werden: die Quelle und den Filter. Das Prinzip ist sehr einfach. Ein Quellsignal durchläuft einen Resonanzfilter.
Die Quelle entspricht dem Anregungssignal u(n), welches der Anregung des menschlichen Vokaltraktes ähnelt. Bei stimmhaften Lauten erzeugen die Stimmbänder eine quasiperiodische Schwingung. Bei stimmlosen Lauten hingegen ensteht an Engstellen im Mund- oder Rachenraum bei geöffneten Stimmbändern eine turbulente Strömung, ein Rauschsignal. Die quasiperiodische Schwingung und das Rauschen besitzen ein flaches Spektrum und dienen als Anregungssignale für die

Spracherzeugung [2].Folglich sind sie die Eingangssignale für den rein rekursiven, auto- regressiven (AR) digitalen Filter H(z), der die Resonanzen des Vokaltraktes modelliert. Sämtliche Modellparameter, wie Grundfrequenz, stimmhaft/stimmlos-Unterscheidung, Gain und Resonanzen sind stark zeitvariant, da sich beim Sprechen das Anregungssignal und die Eigenschaften des Vokaltraktfilters ständig ändern. Wird das zu bearbeitende Sprachsignal in Blöcke von 10 bis 30 ms unterteilt, kann man das System als kurzzeitstationär betrachten.

2.2 Lineare Prädiktion

Geht man also von dem oben genannten Quelle- Filter Modell der Spracherzeugung aus, ist es wünschenswert die Koeffizienten des diesem Modell zugrundeliegenden Allpol- Filters zu bestimmen. Dies ist mit Hilfe der linearen Prädiktion im Sinne einer Systemidentifikation möglich. Das vorhandene Sprachsignal lässt sich somit in die Bestandteile Quelle und Filter, bzw. Anregung und spektrale Einhüllende zerlegen.
Da die „wahren“ Modellkoeffizienten c_k unbekannt sind, kann bei der linearen Prädiktion (LPC) das jeweils aktuelle Sample des Ausgangssignals $s(n)$ aus einer Linearkombination vergangener Samples $s(n-i)(i=1,2,...k)$ und der Verwendung eines FIR Filters mit Koeffizienten a_k angenähert werden.

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{2.3}$$

p bezeichnet die LPC Ordnung und a_k die zu bestimmenden Prädiktionskoeffizienten. Die Differenz zwischen Originalsignal $s(n)$ und geschätztem Signal $\hat{s}(n)$ lautet unter der Voraussetzung $a_0 = 1$:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) = \sum_{k=0}^{p} a_k s(n-k). \tag{2.4}$$

Das Differenzsignal $e(n)$ wird als Residuum oder Prädiktionsfehler bezeichnet, der minimiert werden soll. Bestimmt man die Prädiktorkoeffizienten a_k so, dass sie den „wahren“ Modellkoeffizienten c_k entsprechen, dann ist der Prädiktionsfehler gleich der gesuchten Anregung $u(n)$.
Im Frequenzbereich ergibt sich die Übertragungsfunktion des FIR Prädiktionsfehlerfilters, oder auch LP Analyse Filter genannt, folgendermassen:

$$A(z) = \sum_{k=0}^{p} a_k z^{-k}. \tag{2.5}$$

Damit schreibt sich Gleichung 2.2 im Frequenzbereich

$$E(z) = S(z)A(z). \tag{2.6}$$

Mit der Definition des Allpol- oder auch LP- Synthese- Filters H(z)

$$H(z) = \frac{1}{A(z)} \tag{2.7}$$

erhält man das Eingangssignal S(z) durch Filterung der Anregung U(z) mit dem Synthese Filter H(z).

$$S(z) = U(z) \cdot H(z). \tag{2.8}$$

H(z) modelliert die Resonanzen der Sprache, entspricht also der spektralen Einhüllenden des Signals.
Der Synthese Filter ist minimalphasig, weshalb auch immer eine stabile Inverse A(z) existiert.
Führt man nun eine LPC Analyse des Eingangssprachsignals durch, dann erhält man die spektrale Einhüllende des Signals in Form der AR- Koeffizienten a_k. Die AR Koeffizienten lassen sich mit der Autokorrelationsmethode und dem Levinson Durbin Algorithmus bestimmen.
Da die LPC Analyse nur mit einer endlichen Ordnung p durchgeführt wird, erhalten wir ein geglättetes Spektrum H(z).Mit der Inverse von H(z), dem Analyse Filter A(z) lässt sich nun eine Schätzung des Anregungssignals u(n) ermitteln.

2.3 Bandbreitenerweiterung

Um von einem schmalbandigen Sprachsignal, welches über 4 kHz keine Frequenzen mehr enthält, die originale akustische Bandbreite zu rekonstruieren, braucht man zusätzliches Wissen über das Eingangssignal. Dieses Wissen steht uns, wie in Kapitel 2.1 beschrieben, in Form eines mathematischen Modells zur Verfügung, dem sogenannten Quelle- Filter Modell. Mit Hilfe dieses Modells wird die Entstehung jedes Sprachsignals, ob Schmalband oder Breitband, auf die selben Grundparameter zurückgeführt. Das heisst, das Anregungssignal durchläuft einen Vokaltraktfilter, der durch autoregressive Koeffizienten beschrieben werden kann. Erst dadurch ist es möglich, aus schmalbandigen Sprachsignalen die zugehörige Breitbandsprache zu rekonstruieren. Da das Modell nur eine Schätzung der realen physiologischen Sprachentstehung darstellt, kommt es zu Schätzfehlern, die die Signalqualität der rekonstruierten Sprache beeinträchtigen.
Der in diesem Projekt implementierte Bandbreitenerweiterungs- Algorithmus beruht auf dem vereinfachten Quelle- Filter Modell der Spracherzeugung und ist deshalb auch nur für Sprachsignale geeignet. Die Bandbreitenerweiterung lässt sich

entsprechend des Modells in zwei Schritte unterteilen, erstens der Erweiterung des Anregungssignals und zweitens der Erweiterung der spektralen Einhüllenden. Wie in [4, S. 180] gezeigt, ist die Schätzung der spektralen Einhüllenden für die Qualität der erzeugten Breitbandsprache weitaus wichtiger, als die Schätzung der Anregung.
Abbildung 2.2 zeigt das Blockdiagramm eines BWE Systems in allgemeiner Form.

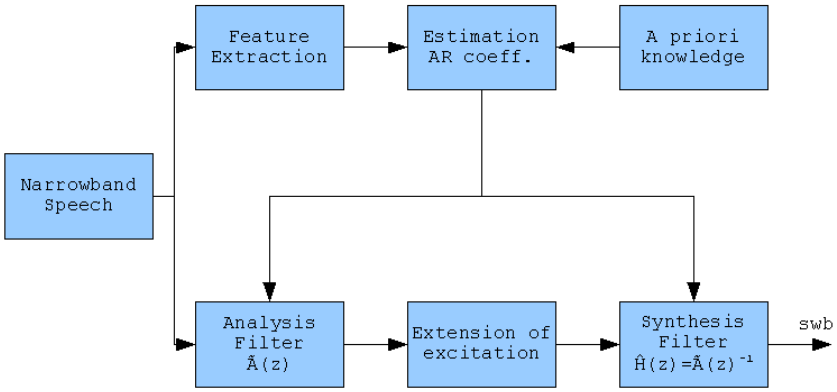

Abbildung 2.2: Blockdiagramm eines BWE Systems

2.3.1 Schmalbandige Sprache

In digitalen Sprachübertragungssystemen wird das Sprachsignal vor der AD Wandlung bandpassgefiltert (Norm ITU-T G.120) und dann mit einer Samplingrate von 8 kHz über den Telefonkanal übertragen. Allgemein besitzt schmalbandige Sprache Frequenzinformationen im Bereich von 300 Hz bis 3400 kHz. Für die BWE des Signals bis zu 8 kHz muss die Samplingrate entsprechend auf 16 kHz erhöht werden. Das kann durch Einfügen von Nullstellen im Zeitbereich und anschliessender Tiefpassfilterung geschehen. Das schmalbandige Signal bleibt auch nach dem Upsampling schmalbandig.

2.3.2 Merkmalsbestimmung und Schätzen der AR- Koeffizienten

Vom schmalbandigen Signal wird mittels Algorithmen zur Merkmalsbestimmung (*Feature Extraction*) möglichst viel Information über das Signal gesammelt. Diese

nutzt man, um mit Hilfe von vorher durch Training ermittelten Transformationen die breitbandigen AR- Koeffizienten aus den schmalbandigen AR- Koeffizienten zu schätzen. Die genaue Verfahrensweise des Trainings und der Schätzung wird in Kapitel 3.2.3 beschrieben. Die neuen AR- Koeffizienten repräsentieren die spektrale Einhüllende des breitbandigen Signals.

2.3.3 Analyse Filter

Die ermittelten breitbandigen AR Koeffizienten $\tilde{a}_k$ fungieren nun als Filterkoeffizienten für den FIR Analyse Filter $\tilde{A}(z)$ an dessen Eingang das schmalbandige Sprachsignal anliegt

$$\tilde{A}(z) = \sum_{k=0}^{p} \tilde{a}_k z^{-k} \tag{2.9}$$

und

$$\tilde{u}_{nb}(n) = \sum_{k=0}^{p} \tilde{a}_k s_{nb}(n-k). \tag{2.10}$$

p bezeichnet die LPC Ordnung und somit die Anzahl der AR- Koeffizienten. Da der Analyse Filter die Inverse des oben behandelten Vokaltraktfilters ist, kann der Ausgang des Analyse Filters $\tilde{u}_{nb}(n)$ als eine Schätzung des immer noch schmalbandigen Anregungssignals gesehen werden.

2.3.4 Erweiterung des Anregungssignals

Das Subsystem der Erweiterung des Anregungssignals befindet sich im unteren Bereich von Abbildung 2.2. Aufgabe dieses Subsystems ist die Wiederherstellung der spektralen Feinstruktur des Sprachsignals. Als Eingang dient die geschätzte schmalbandige Anregung $\tilde{u}_{nb}(n)$. Wichtig ist, dass die Anregung vom schmalbandigen Signal im erweiterten Anregungssignal unverändert erhalten bleibt. Für die Erweiterung des Anregungssignals gibt es viele verschiedene Methoden, die entweder explizit neue Signalkomponenten erzeugen (Nichtlineare Verzerrung wie Vollweggleichrichter, Rauschen, Rauschen und/oder Sinus) oder Komponenten der vorhandenen bandbegrenzten Anregung wiederverwenden (Modulation, Spectral Translation, Pitch Scaling) (siehe [4, S. 185 ff.]). In diesem Projekt wird jeweils eine Methode aus einer dieser beiden Gruppen für die Erweiterung des Anregungssignals verwendet, der Vollweggleichrichter und die Modulation im Zeitbereich.

2.3.5 Synthese Filter

Das bandbreitenerweiterte Anregungssignal $\tilde{u}_{wb}(n)$ dient wiederum als Eingang für den folgenden Synthesefilter, wo schließlich die erweiterte Anregung mit der erweiterten spektralen Einhüllenden kombiniert wird.

$$\tilde{H}(z) = \left(\sum_{k=0}^{p} \tilde{a}_k z^{-k} \right)^{-1} = \frac{1}{\tilde{A}(z)}. \tag{2.11}$$

Unter Berücksichtigung der Normalisierung der AR Koeffizienten ($\tilde{a}_0 = 1$) lässt sich der Ausgang des gesamten BWE Systems wie folgt berechnen

$$\tilde{s}_{wb}(n) = \tilde{u}_{wb}(n) - \sum_{k=1}^{p} \tilde{a}_k \tilde{s}_{wb}(n-k). \tag{2.12}$$

Die Übertragungsfunktionen von Analyse- und Synthese- Filter sind invers zueinander, da bei beiden Filtern die gleichen AR- Koeffizienten benutzt werden.

Kapitel 3

Implementierung des BWE Systems

Das im Folgenden beschriebene System zur Bandbreitenerweiterung von schmalbandigen Sprachsignalen wurde in MATLAB implementiert.
Als Eingangssignal dient ein *Test-Set* der TIMIT Sprachdatenbank (näheres unter Kapitel 4).
Das Eingangssignal wird zuerst Hanning gefenstert mit einer Fensterlänge von 256 Samples und einer *hop size* von 128 Samples. 256 Samples entsprechen 16 ms bei einer Samplingrate von 16 kHz. Damit wird sichergestellt, dass das Signal als quasistationär angesehen werden kann. Die gesamte Signalverarbeitung erfolgt nun blockweise.

3.1 Erweiterung des Anregungssignals

3.1.1 Vollweggleichrichter

Ein sehr einfaches und recheneffizientes Verfahren zur Erweiterung des Anregungssignals ist die Verwendung eines Vollweggleichrichters (*fullwave rectifier*). Dabei wird vom Eingangsignal der Absolutbetrag genommen. Dadurch entstehen zusätzliche geradzahlige Harmonische. Die Grundfrequenz wird unterdrückt, siehe Abbildung 3.1. Bei der Erweiterung der Anregung muss gewährleistet sein, dass

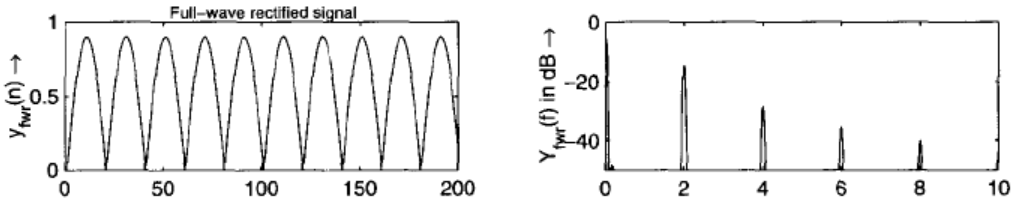

Abbildung 3.1: Vollweggleichrichter [5, S. 127]

das schmalbandige Anregungssignal unverändert erhalten bleibt. Dazu wird das gleichgerichtete Signal hochpassgefiltert mit einer *Cut off*- Frequenz von 3400 Hz und dann zur schmalbandigen Anregung addiert. Als Resultat erhält man das erweiterte Anregungssignal.

3.1.2 Modulation im Zeitbereich

Eine andere Möglichkeit der Erweiterung des Anregungssignals ist die Modulation im Zeitbereich. Durch die Modulation wird das Signal im Frequenzbereich in das gewünschte Frequenzband verschoben. In diesem Projekt wird das Signal mit einer reellwertigen Modulationsfunktion moduliert. In Folge dessen entstehen Überlappungen der verschobenen Spektren, die durch Filterung wieder beseitigt werden müssen. Die Modulation mit einer reellen Cosinusfunktion lässt sich wie folgt darstellen:

$$\tilde{u}_M(k) = \tilde{u}_{nb}(k) \cdot \xi cos(\Omega_M k) \tag{3.1}$$

Der fixe Wert ξ ist abhängig von der jeweiligen Modulationsfrequenz Ω_M und muss aus $\xi \in \{1, 2\}$ gewählt werden, um die korrekte Leistung des erweiterten Anregungssignals zu bestimmen. Durch Multiplikation des Eingangssignals mit dem Cosinus entstehen zwei verschobene Kopien des Originalspektrums $\tilde{U}_{nb}(e^{j\Omega})$.

$$\tilde{U}_M(e^{j\Omega}) = \frac{\xi}{2}(\tilde{U}_{nb}(e^{j(\Omega-\Omega_M)}) + \tilde{U}_{nb}(e^{j(\Omega+\Omega_M)})) \tag{3.2}$$

In diesem Fall wird ein ξ Wert von 1 gewählt und die Modulation zweimal durchgeführt. Bei der ersten Modulation wird das Schmalbandspektrum von 1100 Hz bis 3400 Hz einmal nach oben verschoben, so dass sich die Anregung nach korrekter Filterung von 300 Hz bis 5.7 kHz erstreckte. Bei der zweiten Modulation wird das gleiche Schmalbandspektrum noch weiter in den höheren Frequenzbereich moduliert. Folglich reicht das erweiterte Anregungssignal von 300 Hz bis 8 kHz.

3.2 Erweiterung der spektralen Einhüllenden

Der wichtigste Schritt bei der Bandbreitenerweiterung ist das Schätzen der breitbandigen spektralen Einhüllenden. Die meisten BWE Algorithmen benutzen dazu statistische Schätzmethoden (siehe [4, S.196 ff.]). In dieser Arbeit wird das *Linear Mapping* verwendet.

3.2.1 Merkmalsbestimmung

Von jedem Block des schmalbandigen Eingangssignals werden Merkmale extrahiert, die Informationen über die spektrale Einhüllende enthalten. Dazu wird vom

schmalbandigen und breitbandigen Signal eine LPC Analyse der Ordnung $p_n = 9$ bzw. $p_w = 18$ durchgeführt. Die LPC Koeffizienten können nicht direkt für die weitere Verarbeitung verwendet werden, da sie zu empfindlich gegenüber Quantisierungsfehlern sind und somit die Gefahr von Instabilität besteht. Als Abhilfe werden die LPC Koeffizienten in stabilere Merkmale transfomiert, die *Line Spectral Frequencies* (LSF) und die aus den LPC Koeffizienten berechneten Cepstralen Koeffizienten.

3.2.1.1 Line Spectral Frequencies

Line Spectral Frequencies (LSF´s) bieten eine alternative Darstellungsform der LPC Koeffizienten, die aber einige wichtige Vorteile bringt. Sie sind robust gegenüber Quantisierungsfehlern, Interpolation ist möglich und unter den unten genannten Bedingungen bilden sie ein stabiles System. LSF´s lassen sich berechnen, indem man den LP Analyse Filter A(z) (siehe Gleichung 2.9) der Ordnung p in ein Spiegelpolynom P(z) und in ein Antispiegelpolynom Q(z) zerlegt.

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{3.3}$$
$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{3.4}$$

Die Nullstellen dieser sogenannten *Line Spectral Pair* (LSP)- Polynome bilden die LSF.
Das Polynom A(z) lässt sich daraus wieder eindeutig rekonstruieren.

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{3.5}$$

Betrachten wir nun die Nullstellen α_i und β_i von P(z) und Q(z). Diese Nullstellen befinden sich auf dem Einheitskreis der z- Ebene, $|\alpha_i| = |\beta_i| = 1$ und lassen sich wie folgt darstellen:

$$\alpha_i = e^{i\pi\lambda_i} \text{ und } \beta_i = e^{i\pi\gamma_i}. \tag{3.6}$$

Die LSF Parameter λ_i und γ_i bezeichnen die Winkel der Nullstellen auf dem Einheitskreis. Sie liegen in aufsteigender Reihenfolge im Bereich zwischen 0 und 1. Mit ihnen lässt sich A(z) wieder herstellen. Weiters muss darauf geachtet werden, dass die Nullstellen der Polynome P(z) und Q(z) ineinander verschachtelt alternieren. Die Summe von beiden Polynomen ergibt ein minimalphasiges System und die Stabilität der Allpol Filter $P^{-1}(z)$ und $Q^{-1}(z)$ ist gesichert. An den Polstellen von $P^{-1}(z)$ und $Q^{-1}(z)$ ergeben sich im Frequenzbereich unendlich hohe Werte, die jeweils den Winkeln der Nullstellen entsprechen. Diese sind im Spektrum als vertikale Linien erkennbar. Diese Linien werden als *Line Spectral Frequencies* bezeichnet.

3.2.1.2 LPC- Cepstrum

Die Übertragungsfunktion eines LP Synthesefilters kann durch ein unendlich lange Folge von Cepstralkoeffizienten ausgedrückt werden. Die Cepstralkoeffizienten c_i lassen sich mit einer einfachen rekursiven Formel aus den LPC Koeffizienten a_i berechnen. p ist die LPC Ordnung.

$$c_0 = ln\sigma^2 \tag{3.7}$$

$$c_i = -a_i - \sum_{n=1}^{i-1} \frac{n}{i} c_n a_{i-n} \tag{3.8}$$

$$\text{für } i > 0, \text{ mit } a_i = 0 \text{ für } i > p \tag{3.9}$$

Die cepstrale Darstellung hat den Vorteil einer besseren Dekorrelation der Koeffizienten, was besser für das Modellieren der Wahrscheinlichkeitsdichtefunktion (PDF) p(x) ist ([4]). Auch wenn in diesem Projekt keine PDF modelliert wird, lässt sich doch ein anderer Vorteil der c_i nutzen. Mit ihnen lässt sich die Log Spectral Distortion (LSD) berechnen, ein Fehlermaß, welches weiter unten im Detail beschrieben wird.

$$d_{LSD}^2 = \left(\frac{10}{ln10}\right)^2 \sum_{i=-\infty}^{\infty} (c_i - \tilde{c}_i)^2. \tag{3.10}$$

3.2.2 Erweiterung der AR Koeffizienten

Die Schätzung der erweiterten AR Koeffizienten mittles *Gaussian Mixture Models* ist weit verbreitet, jedoch sehr aufwendig zu berechnen. Echtzeitanwendungen erfordern recheneffizientere Methoden, wie zum Beispiel das *Linear Mapping*.

3.2.3 Linear Mapping

Um die breitbandige spektrale Einhüllende aus einem Merkmalsvektor x des schmalbandigen Eingangssignals zu schätzen, wird in dieser Arbeit die Methode des *Linear Mappings*[4, S. 219 ff.] angewandt. Weitere Methoden sind in [4, S. 217 ff.] beschrieben. Beim *Linear Mapping* wird eine lineare Beziehung zwischen Schmalbandspektrum und Breitbandspektrum vorausgesetzt. Die geschätzten Breitband Features $\tilde{\mathbf{y}}$ erhält man aus dem Schmalband- Merkmalsvektor $\mathbf{x}$ und der Transformationsmatrix $\mathbf{A}$:

$$\tilde{\mathbf{y}} = \mathbf{A}^T \cdot \mathbf{x}. \tag{3.11}$$

Die Matrix $\mathbf{A}$ hat die Dimension von b x d mit $b = dim\ \mathbf{x}$ und $d = dim\ \mathbf{y}$. Die gesamten Informationen über die gegenseitigen Abhängigkeiten von $\mathbf{x}$ und $\mathbf{y}$ sind

in der Transformationsmatrix $\mathbf{A}$ enthalten, welche in einer *off- line* Trainingsphase des BWE Systems erzeugt wird. Das *Linear Mapping* ist sehr einfach zu implementieren, benötigt keinen grossen Speicherplatz und ist sehr recheneffizient. Der große Nachteil ist jedoch, dass es in Wirklichkeit keine einfache lineare Beziehung zwischen Featurevektor $\mathbf{x}$ und der Repräsentation der breitbandigen spektralen Einhüllenden $\mathbf{y}$ gibt. So kann es zu teilweise sehr starken Artefakten im geschätzten Spektrum kommen.

3.2.3.1 Trainingsphase

Ziel des Trainings ist es, die Transformationsmatrix $\mathbf{A}$ zu bestimmen. Dafür wurden mit 16 kHz gesampelte Breitbandsprachproben aus der TIMIT Sprachdatenbank verwendet (näheres unter Kapitel 4). Insgesamt ergab sich ein Trainingsset von knapp 9 Minuten Länge.
Die zugehörige Schmalbandsprache entstand durch Bandpassfilterung des breitbandigen Originals. Der Frequenzgang des verwendeten Bandpassfilters ist in Abbildung 3.2 zu sehen.

Es werden von jedem Block beider Signale die Merkmale $\mathbf{y}$ und $\mathbf{x}$, also die *Line Spectral Frequencies* bzw. die Cepstralen Koeffizienten, bestimmt und in den Matrizen F_y und F_x gespeichert. Das Training muss für beide *Feature Extraction* Methoden getrennt durchgeführt werden.
In den Zeilen der beiden Matrizen F_y und F_x befinden sich die Merkmalsvektoren für Breitband- und die zugehörige Schmalbandsprache. Die optimale Matrix $\mathbf{A}$ entsteht als Ergebnis einer Minimierung des Fehlers $\mathbf{y} - \mathbf{A}^T\mathbf{x}$ für das gesamte Trainingsset. Das entspricht der Methode der kleinsten Fehlerquadrate (*least sqares optimization*).

$$e^2 = tr\left[(\mathbf{F}_y - \mathbf{F}_x\mathbf{A})^T(\mathbf{F}_y - \mathbf{F}_x\mathbf{A})\right]. \tag{3.12}$$

e^2 ist die Summe der Quadrate aller Differenzen $y_i(m) - \tilde{y}_i(m)$, ein Schätzfehler, der minimiert werden soll. Dazu muss die obige Gleichung nach den einzelnen Elementen der Matrix $\mathbf{A}$ a_{ij} abgeleitet und null gesetzt werden. Das führt zu:

$$(\mathbf{F}_y - \mathbf{F}_x\mathbf{A})^T\mathbf{F}_x \equiv 0. \tag{3.13}$$

Da diese Beziehung unabhängig von den Positionen der Elemente a_{ij} ist, lassen sich so alle Elemente der Transformationsmatrix $\mathbf{A}$ bestimmen. Der Trainingsalgorithmus lässt sich wie folgt darstellen:

$$\mathbf{A} = \left(\mathbf{F}_x^T\mathbf{F}_x\right)^{-1}\mathbf{F}_x^T\mathbf{F}_y. \tag{3.14}$$

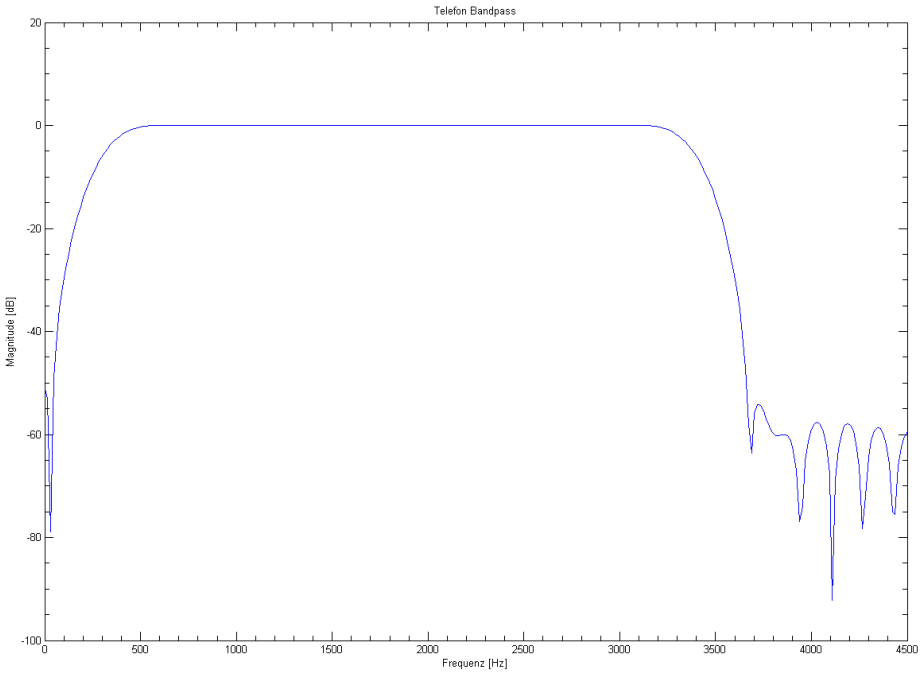

Abbildung 3.2: Verwendeter Telefon Bandpassfilter (nach ITU-T G.120)

3.3 Fehlermaße

Um die Performance des BWE Algorithmus mit verschiedenen Anregungserweiterungen und Merkmalen objektiv einschätzen zu können, ist es notwendig Fehlermaße zu definieren. Im Zug dieses Projekts werden drei gebräuchliche Fehlermaße aus der Sprachsignalverarbeitung verwendet, die *Log Spectral Distortion* (LSD), die *Itakura Saito Distance* (DIS) und die *Cepstral Distance* (CEPD). Mit diesen Fehlermaßen lassen sich die ursprünglichen originalen breitbandigen Sprachsignale mit der geschätzten breitbandigen Sprache vergleichen.
Die ***Log Spectral Distortion***, der gemittelte Unterschied zwischen zwei Leistungsdichtespektren P und $\hat{P}$ schreibt sich wie folgt:

$$d_{LSD}(P,\hat{P}) = \frac{1}{\Delta F}\int_{\Delta F}\left|10\,log_{10}P(f) - 10\,log_{10}\hat{P}(f)\right| df. \tag{3.15}$$

Die LSD wird in dB angegeben.

Die ***Itakura Saito Distance*** ist ein asymmetrisches Fehlermaß, welches eine bessere Aussage über die subjektiv wahrgenommene Sprachqualität als die LSD

erlaubt. Werte entsprechend $P(f) > \hat{P}(f)$ werden stärker bestraft als $\hat{P}(f) > P(f)$. Das heisst, perzeptiv wichtige Informationen, wie Formanten werden stärker gewichtet [6].

$$d_{IS}(P, \hat{P}) = \frac{1}{\Delta F} \int_{\Delta F} \left[\frac{P(f)}{\hat{P}(f)} - ln \left(\frac{P(f)}{\hat{P}(f)} \right) - 1 \right] df. \tag{3.16}$$

Die ***Cepstral Distance*** wird oft in der Sprachsignalverarbeitung als Fehlermaß verwendet. Sie ist der LSD ähnlich, man erspart sich aber die Berechnung des Logarithmus und der FFT. [3, S. 34]

$$d_{CD}(c, \hat{c}) = \sum_{i=1}^{p} (c - \hat{c}_i)^2. \tag{3.17}$$

Kapitel 4

Experimente

In diesem Kapitel werden die durchgeführten Experimente und deren Ergebnisse beschrieben.

4.1 TIMIT Datenbank

Für das Training und die Test- Experimente wurde die TIMIT Datenbank verwendet. Sie ist unter anderem vom Massachusetts Institute of Technology (MIT) und Texas Instruments (TI) entwickelt worden und enthält Sprachaufnahmen von 630 SprecherInnen mit 8 verschiedenen Dialekten im amerikanischen Englisch, wobei jeder Sprecher 10 phonetisch reichhaltige Sätze liest. Die Aufnahmen sind mit einer Samplingfrequenz von 16 kHz abgetastet worden. Die verwendeten Sätze bestehen aus 3 phonetischen Gruppen:
1. Dialekt- Sätze (SA) sollen die Eigenschaften des Dialekts hervorheben.
2. Phonetisch- kompakte Sätze (SX) enthalten Phoneme, die entweder schwierig oder von besonderem Interesse sind.
3. Phonetisch- variable Sätze (SI) wurden aus Büchern ausgewählt, um die Variabilität des allophonischen Zusammenhangs zu erhöhen. [3, S.43/44]

Für das Training wurden 30 Sprecher (15 Frauen, 15 Männer) und für die Testphase 2 Sprecher (1 Frau, 1 Mann) der Dialog- Region 3 (North Midland) verwendet. Jeder Sprecher der Trainingsphase liest 3 Sätze aus der SI und 3 aus der SX Gruppe, wobei die SprecherInnen der Testphase von denen der Trainingsphase verschieden sind.
Damit die Qualität der linearen Transformation verbessert wird, werden zwei Trainingsmatrizen erstellt, eine für stimmhafte und eine für stimmlose Blöcke. Die stimmhaft/ stimmlos- Unterscheidung wird durch einen Nulldurchgangszähler implementiert.

4.2 Ergebnisse

In Abbildung 4.1 sind die berechneten Fehler der verschiedenen Methoden aufgelistet. Die zugehörigen Spektrogramme befinden sich im Anhang.

Excitation Mode	Spectral Estimation	Speech Signal	Distortion Measures lsd [dB]	dis	cepd
VWG	LSF	fm	**4,8617**	**8,9716**	**0,12097**
Modulation	LSF	fm	**4,8755**	**8,8133**	**0,12097**
VWG	Cepstral coeff.	fm	**7,1337**	**10,18**	**0,04726**
Modulation	Cepstral coeff.	fm	**7,5723**	**7,4699**	**0,04726**

Abbildung 4.1: Fehlermaße

Insgesamt ergeben sich für die *Line Spectral Frequencies* in Verbindung mit dem Vollweggleichrichter bzw. mit der Modulation die besten Ergebnisse. Auch vom Höreindruck her sind diese Features den cepstralen Koeffizienten vorzuziehen. Auch wenn die *Itakura Saito Distance* ein perzeptives Fehlermaß darstellt, kann von ihr nicht direkt auf die empfundene Sprachqualität geschlossen werden. So ist der geringe DIS Wert von cepstralen Koeffizienten und Modulation in diesem Fall kein Indiz für gute Sprachqualität. Es können starke Artefakte auftreten, auch wenn die DIS relativ gering ausfällt. Deshalb ist es auf jeden Fall wichtig, mehrere Fehlermaße zu verwenden und miteinander zu vergleichen.
Die viel niedrigeren Fehlerwerte der cepstralen Distanz für die cepstralen Koeffizienten im Vergleich zu den LSF sind darauf zurückzuführen, dass beim Training mit den cepstralen Koeffizienten genau diese cepstrale Distanz minimiert wird. Im Zuge dieser Arbeit wurde kein Hörtest durchgeführt. Beim persönlichen Hörvergleich zwischen schmalbandiger und erweiterter Sprache lassen sich eine etwas verbesserte Verständlichkeit, der Frequenzbereich nach oben ist erweitert, aber auch Verzerrungen, die den subjektiven Allgemeineindruck stark mindern, feststellen. Dies ist in erster Linie auf das lineare Mapping zurückzuführen, welches ein lineare Beziehung zwischen unterem und oberem Frequenzband attestiert.

Kapitel 5

Ausblick

In diesem Kapitel werden einige Vorschläge zur Verbesserung der Qualität des Bandbreitenerweiterungs-Algorithmus gemacht.
Das primäre Problem bei der Bandbreitenerweiterung ist die Erweiterung der spektralen Einhüllenden. Hier geht die meiste Performance verloren. Da in dieser Arbeit das lineare Mapping für die Transformation zwischen schmalbandiger und breitbandiger Sprache verwendet wird, könnte man dort ansetzen und andere Arten von Transformationen einführen, wie zum Beispiel das *Codebook Mapping* oder das Verwenden von statistischen Modellen [7, S. 8 ff.]. Dadurch lässt sich die Sprachqualität weiter verbessern. Unter Beibehaltung des linearen Mappings gibt es weitere Möglichkeiten der Optimierung. Die Modellierung der spektralen Einhüllenden kann über andere Merkmale erfolgen, wie zum Beispiel *Mel-Frequency Cepstral Coefficients* oder cepstrale Koeffizienten.
Weiters lässt sich das erweiterte Anregungssignal mit aufwendigeren Methoden (pitch- adaptive Modulation, [4, S. 192]) genauer modellieren.
Für die Erstellung der beiden Trainingsmatrizen *A- stimmhaft* und *A- stimmlos* ist es notwendig, das Signal in stimmhafte und stimmlose Blöcke einzuteilen. Dies lässt sich eventuell mit Merkmalen, wie Autokorrelationskoeffizienten und Gradientenindex genauer tun, als mit dem Zählen der Nulldurchgänge.
Nachdem man das erweiterte Sprachsignal geschätzt hat, ist eine Nachbearbeitung notwendig. So kann, je nachdem ob ein Fehlermaß einen bestimmten Schwellwert überschreitet, wieder zum schmalbandigen Signal zurückgekehrt werden.
Aber auch wenn der Fehler unter der Schwelle bleibt, können starke Artefakte auftreten. Deshalb ist es besser, sich die Standardabweichung von Schmalband- und erweitertem Signal anzuschauen und diese als Kriterium zu nehmen.[3, S. 36]
In den meisten bestehenden BWE Systemen wird die bandbreitenerweiterte Sprache der schmalbandigen Sprache vorgezogen. Trotzdem ist die Sprachqualität auch bei den *state of the art* Systemen mit Gaußschen Mischmodellen und Hidden Markow Modellen immer noch um einiges schlechter als die originale Breitbandsprache.

Dieser Unterschied liegt im Schnitt bei einer Log Spectral Distortion von 3.2 dB für hohe Frequenzen (3.4 - 7 kHz) und 2.3 dB für tiefe Frequenzen (50- 300 Hz). Ob sich Dieser Unterschied lässt sich sicher nicht ohne sehr großen Aufwand verkleinern.
Bisher wurden BWE Systeme hauptsächlich für saubere, rauschfreie Sprachsignale entwickelt. Ziel weiterführender Entwicklung sollte es sein, die Algorithmen robust gegenüber Rauschen und Verzerrungen des schmalbandigen Signals zu machen oder auch andere Signale wie Musik zuzulassen [4, S. 235].

Anhang A

Ergebnisplots

Die folgenden Abbildungen enthalten die Spektrogramme der schmalbandigen, der originalen breitbandigen und der bandbreitenerweiterten Testsignale. VWG steht für Vollweggleichrichter, Mod für Modulation, LSF für *Line Spectral Frequencies* und Cep für die aus den LPC Koeffizienten abgeleiteten cepstralen Koeffizienten. Es wurde jeweils ein Satz von einer Frauenstimme und einer Männerstimme ausgewählt.
Frauenstimme: „*Several firms are merchandising enzyme preparation through feed manufacturers.*“ (TIMIT, DR3, SI 824)
Männerstimme: „*We can, however, maximize its expected value.*“ (TIMIT, DR3, SI 463)

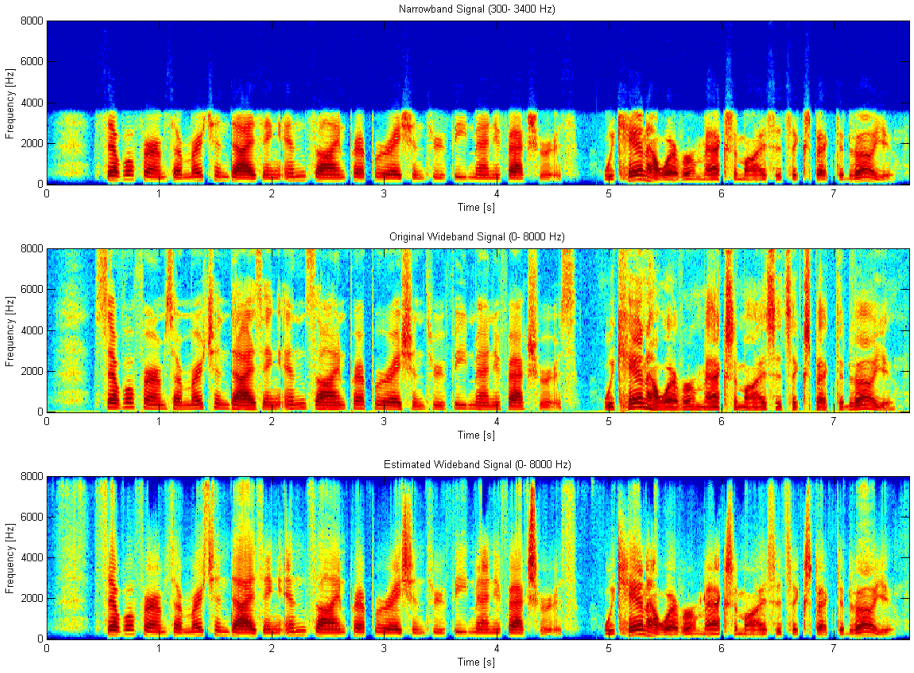

Abbildung A.1: Spektogramme vom Schmalband-, Breitband- und Erweiterten-Signal, (VWG, LSF)

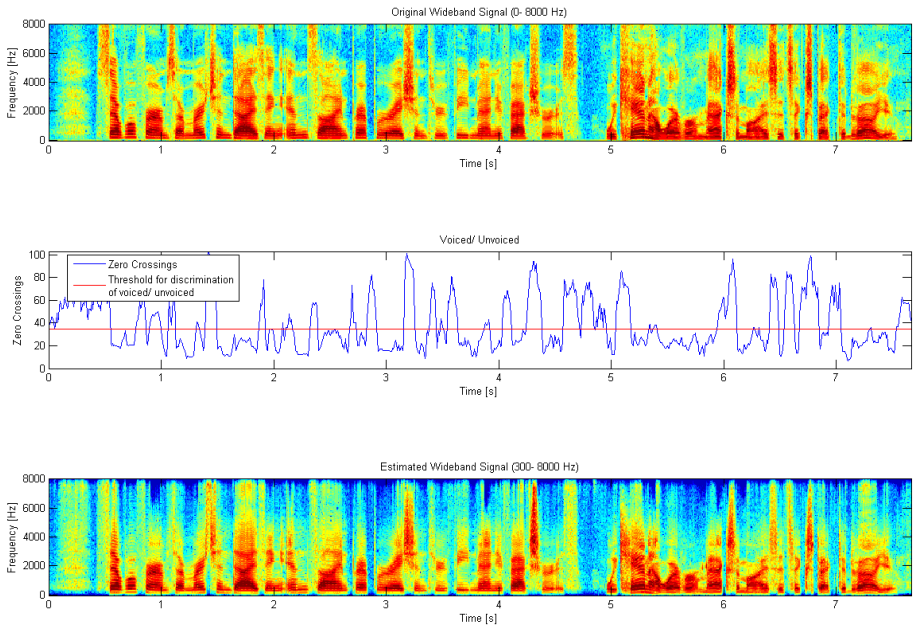

Abbildung A.2: Spektogramme vom Breitband- Signal, von den Nullstellendurchgängen und vom Erweiterten- Signal, (VWG, LSF)

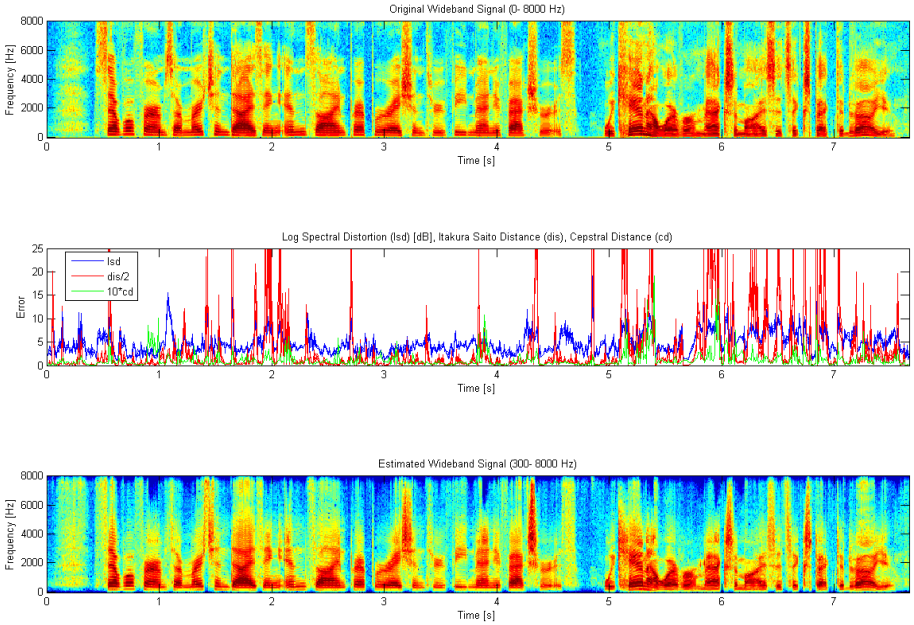

Abbildung A.3: Spektogramme vom Breitband- Signal, den Fehlermaßen und vom Erweiterten- Signal, (VWG, LSF)

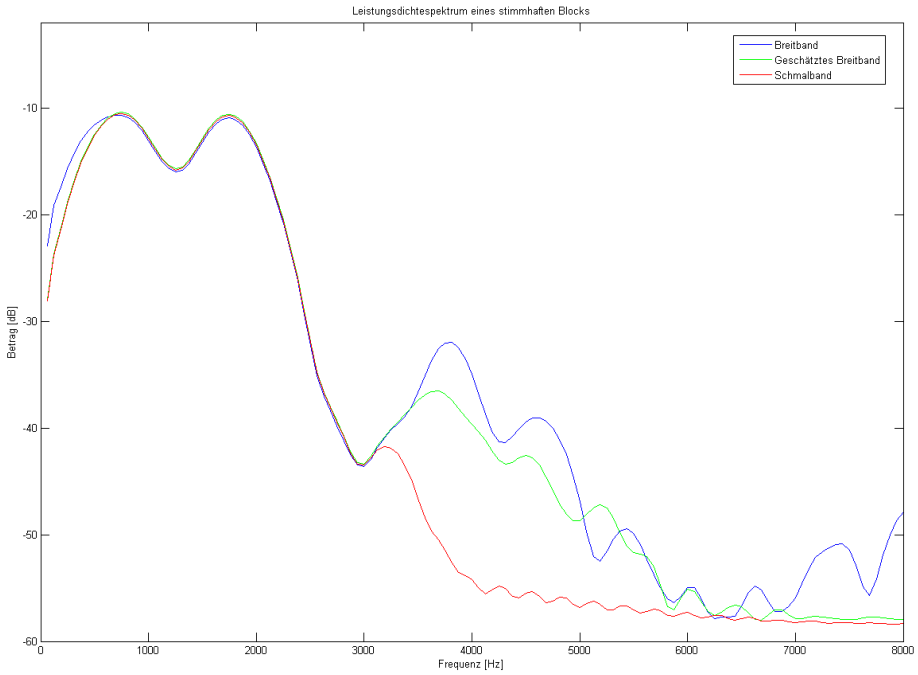

Abbildung A.4: Leistungsdichtespektrum eines 256 Sample langen Blocks eines stimmhaften Lautes, (VWG, LSF)

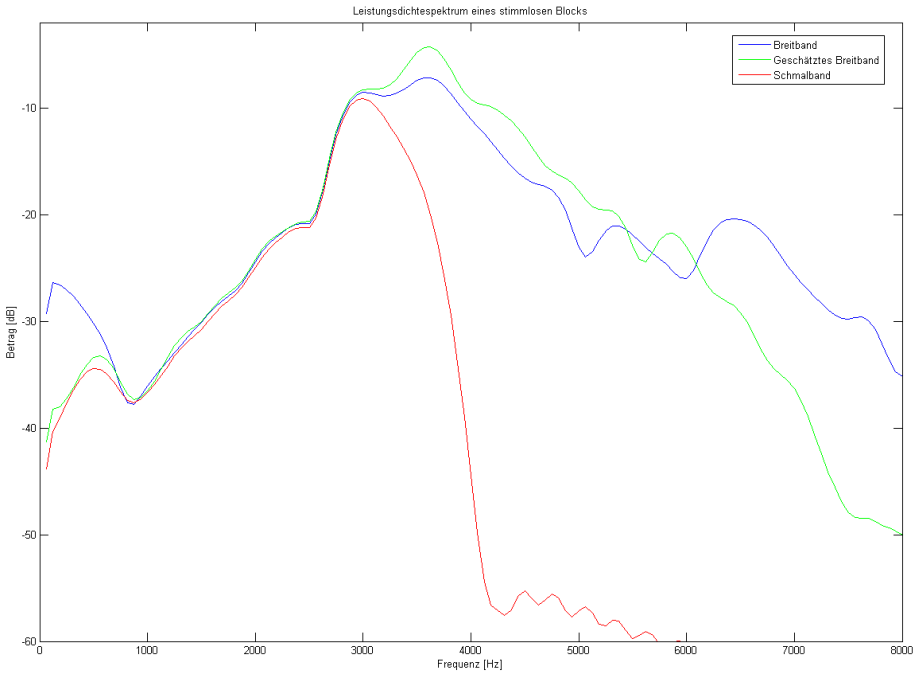

Abbildung A.5: Leistungsdichtespektrum eines 256 Sample langen Blocks eines stimmlosen Lautes, (VWG, LSF)

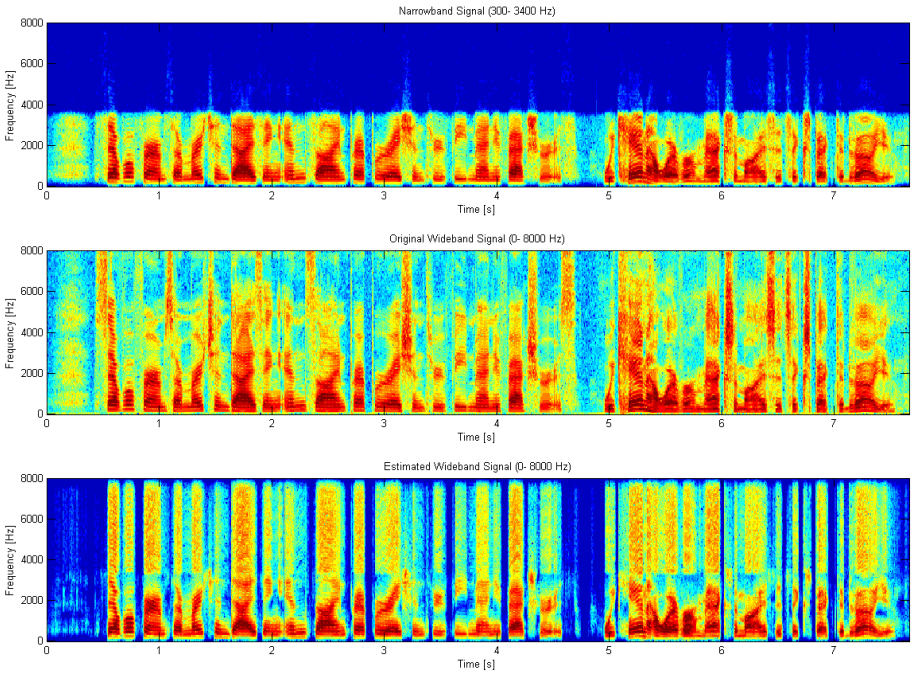

Abbildung A.6: Spektogramme vom Schmalband-, Breitband- und Erweiterten-Signal , (Mod, LSF)

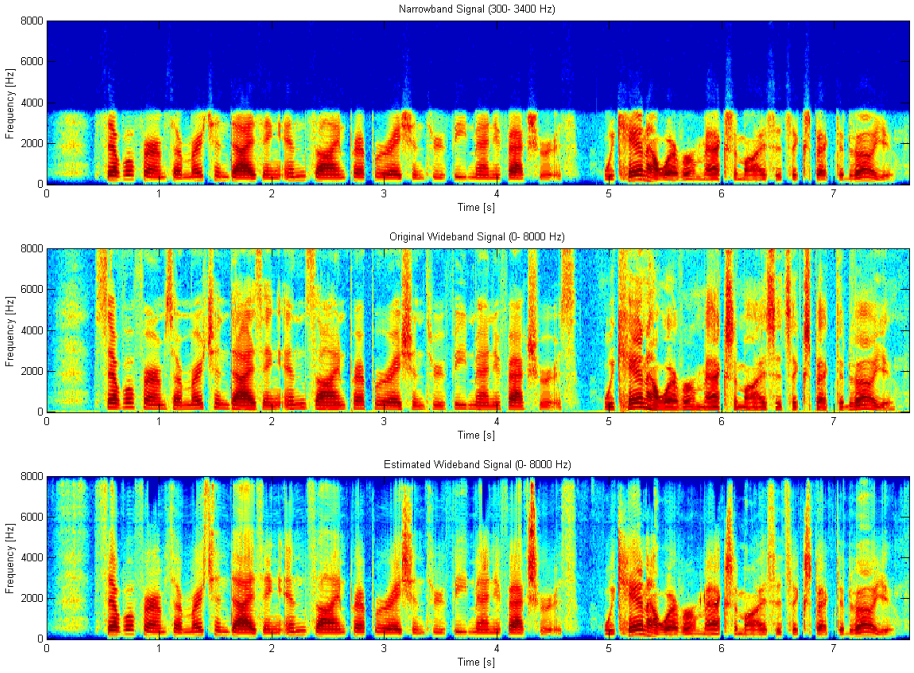

Abbildung A.7: Spektogramme vom Schmalband-, Breitband- und Erweiterten-Signal, (VWG, Cep)

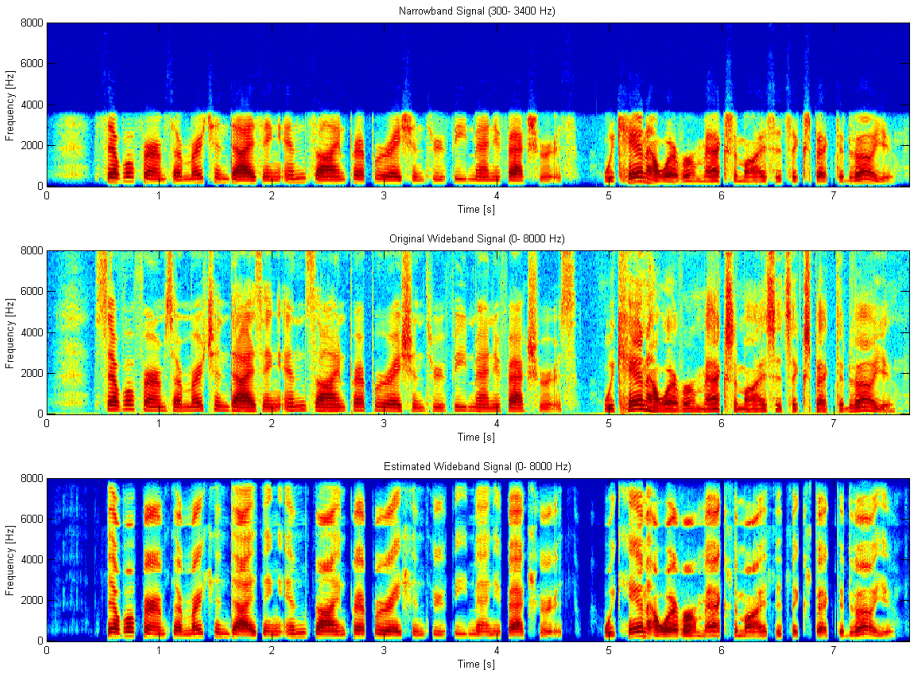

Abbildung A.8: Spektogramme vom Schmalband-, Breitband- und Erweiterten-Signal, (Mod, Cep)

Literaturverzeichnis

[1] M. Nilsson, S. V. Andersen, and W. B. Kleijn, "On the mutual information between frequency bands in speech," *IEEE*, 2000.

[2] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung.* Teubner Verlag Stuttgart, 1998.

[3] D. Zaykovskiy, "On the use of neural networks for vocal tract transfer function estimation," Master's thesis, Information Technology University of Ulm, November 2004.

[4] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension.* John Wiley & Sons, Ltd, 2004.

[5] U. Zölzer, Amatriain, and Arfib..., *DAFX- Digital Audio Effects*, U. Zölzer, Ed. John Wiley & Sons, LTD, 2002.

[6] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Kialt, "Comparative study of several distortion measures for speech recognition," *IEEE*, 1985.

[7] W.-S. Hsu, "Robust bandwidth extension of narrowband speech," Master's thesis, Department of Electrical & Computer Engineering McGill University Montreal, Canada, November 2004.

[8] Y. Arima and T. Shimamura, "Noise-robust speech analysis using system identification methods," *Electronics and Communications in Japan, Part 3*, vol. Vol. 86, No. 3, 2003.

[9] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, vol. VOL. ASSP-24, NO. 3, JUNE 1976.

[10] D. Bansal, B. Raj, and P. Smaragdis, "Bandwidth expansion of narrowband speech using non-negative matrix factorization," December 2005.

[11] G. Chen and V. Parsa, "Hmm-based frequency bandwidth extension for speech enhancement using line spectral frequencies," *IEEE ICASSP*, 2004.

[12] S. Chen and H. Leung, "Artificial bandwidth extension of telephony speech by data hiding," *IEEE*, 2005.

[13] S. Chennoukh, A. Gerrats, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," *IEEE*, 2001.

[14] R. V. Cox, D. Malah, and D. Kapilow, "Improving upon toll quality speech for voip," *IEEE*, 2004.

[15] J. Epps, "Wideband extension of narrowband speech for enhancement and coding," Ph.D. dissertation, School of Electrical Engineering and Telecommunications, University of New South Wales, 2000.

[16] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," *IEEE*, 1999.

[17] J. Epps and Holmes, "Speech enhancement using stc- based bandwidth extension," *Department of Telecommunications, School of Electrical Engineering The University of New South Wales 2052 Australia*, 1998.

[18] B. Geiser, P. Jax, and P. Vary, "Robust wideband enhancement of speech by combined coding and artificial bandwidth extension."

[19] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 2006.

[20] H. Gustafsson, I. Claesson, and U. Lindgren, "Speech bandwidth extension," *IEEE*, 2001.

[21] H. Gustafsson, U. A. Lindgren, and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. VOL. 14, NO. 2, MARCH 2006.

[22] S. Jaisimha and I. Y. Soon, "Bandwidth extension of narrow band speech using cepstral linear prediction," *IEEE*, 2003.

[23] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," *ICASSP*, 2004.

[24] Jax and Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model," *IEEE*, 2003.

[25] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," *IEEE*, 2002.

[26] Jax and Vary, "Wideband extension of telephone speech using a hidden markov model," *IEEE*, 2000.

[27] P. Kabal and B. Kleijn, "All-pole modelling of mixed excitation signals," *IEEE*, 2001.

[28] E. Larsen, R. M. Aarts, and M. Danessis, "Efficient high-frequency bandwidth extension of music and speech," *AES Convention Paper*, vol. 5627, 2002.

[29] L. Liao and M. A. Gregory, "Algorithms for speech classification," *Fifth International Symposium on Signal Processing and its Applications, ISSPA '99, Brisbane, Australia, 22-25 August, 1999*, 1999.

[30] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. VOL. 13, NO. 5, SEPTEMBER 2005.

[31] J. MAKHOUL, "Correction to "linear prediction: A tutorial review"."

[32] J. Makhoul, "Linear prediction: A tutorial review," *PROCEEDINGS OF THE IEEE*, vol. VOL. 63, NO. 4, APRIL 1975.

[33] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech based on linear mapping," *Electronics and Communications in Japan, Part 2*, vol. Vol. 85, No. 8, 2002.

[34] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," *IEEE*, 2001.

[35] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," *IEEE*, 2002.

[36] P. Noll, "Wideband speech and audio coding," *IEEE Communications Magazine*, November 1993.

[37] ——, "Digital audio coding for visual communications," *PROCEEDINGS OF THE IEEE*, vol. VOL. 83, NO. 6, JUNE 1995.

[38] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," *ICASSP IEEE*, 2004.

[39] I. Y. Soon, C. Y. S.N. Koh, and W. Ngo, "Transformation of narrowband speech into wideband speech with aid of zero crossings -rate," *ELECTRONICS LETTERS*, vol. 38 No.24, November, 2002.

[40] I. Y. Soon and C. K. Yeo, "Bandwidth extension of narrowband speech using cepstral analysis," *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, October 2004.

[41] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. Vol. 82 No. 10, October1994.

[42] B. D. Storey, "Computing fourier series and power spectrum with matlab."

[43] K. Tokuda, T. Kobayashi, and S. Imai, "Recursive calculation of mel-cepstrum from lp coefficients," 1994.

[44] B. Wei and J. D. Gibson, "Comparison of distance measures in discrete spectral modeling."

[45] S. Yao and C.-F. Chan, "Block-based bandwidth extension of narrowband speech signal by using cdhmm," *ICASSP IEEE*, 2005.

[46] H. Yasukawa, "Wideband speech recorvery from bandlimited speech in telephone communications," *IEEE*, 1998.

[47] B. Zaykovskiy, Dmitry Iser, "Comparison of neural networks and linear mapping in an application for bandwidth extension," 2005.

[48] Q. Zhao, T. Shimamura, and J. Suzuki, "Improvement of lpc analysis of speech by noise compensation," *Electronics and Communications in Japan, Part 3*, vol. Vol. 83, No. 9, 2000.

[49] H. Özera, I. Avcibas, B. Sankura, and N. Memonc, "Steganalysis of audio based on audio quality metrics."