# Artificial Bandwidth Extension of Band Limited Speech Based on Vocal Tract Shape Estimation

Itai Katsir

# Artificial Bandwidth Extension of Band Limited Speech Based on Vocal Tract Shape Estimation

Final Paper

As Partial Fulfillment of the Requirements for
the Degree of Master of Science in Electrical Engineering

Itai Katsir

Submitted to the Senate of the Technion—Israel Institute of Technology
Kislev 5772 Haifa December 2011

# Acknowledgement

The final paper was done under the supervision of Professor David Malah and Professor Israel Cohen in the Department of Electrical Engineering.
I would like to express my deep gratitude to both of my supervisors, for sharing with me their vast knowledge and spending long hours guiding me throughout all the stages of this research. I appreciate it profoundly.

Many thanks to Nimrod Peleg from the Signal and Image Processing Lab (SIPL).
Special thanks to Prof. Ilan Shallom for sharing with me his knowledge in speech processing.

Finally, I owe my loving thanks to my family - my parents for their help and support, my daughter, Yuval, and my wife, Hallel, for all her encouragement and patience. To them I dedicate this work.

The Generous Financial Help Of The Technion Is Gratefully Acknowledged. This research was supported by the Israel Science Foundation (grant no. 1130/11).

# Contents

# List of Figures

# List of Tables

# Abstract

This research addresses the challenge of improving degraded telephone narrowband speech quality caused by signal band limitation to the range of 0.3 - 3.4 kHz. We introduce a new speech bandwidth extension (BWE) algorithm which estimates and produces the high-band spectral components ranging from 3.4 kHz to 7 kHz, and emphasizes the lower spectral components around 300 Hz.

Using a speech production model known as the source-filter model, the high-band production is separated into two independent algorithms for spectral envelope estimation and for excitation generation. The excitation is generated using a simple spectral copying technique. The spectral envelope is estimated using a statistical approach. It involves phonetic and speaker dependent estimation of the spectral envelope. Speech phoneme information is extracted by using a Hidden Markov Model (HMM). Speaker vocal-tract shape information, corresponding to the wideband signal, is extracted by a codebook search. The proposed method provides better estimation of high-band formant frequencies, especially for voiced sounds, as well as improved estimation of spectral envelope gain, especially for unvoiced sounds. Further processing of the estimated vocal tract shape, including vocal tract shape iterative tuning, reduces artifacts in cases of erroneous estimation of speech phoneme or vocal tract shape.

The low-band is emphasized using an equalizer filter, which improves speech naturalness, especially for voiced sounds.

We present objective experimental results that demonstrate improved wideband quality for different speech sounds in comparison to other BWE methods. Subjective experimental results show improved speech quality of the BWE speech signal compared to the received narrowband speech signals.

# Notation

| | |
|---|---|
| $A_{n_A}$ | Area coefficient of acoustic tube with index $n_A$ |
| $\mathbf{A}_{\mathbf{NB}}$ | NB VTAF coefficients |
| $\mathbf{A}_{\mathbf{WB}}$ | WB VTAF coefficients |
| $f_{n_f}$ | Formant frequency with index $n_f$ |
| $f_s$ | Speech sampling rate |
| $\phi_{\mathrm{WB}}\left(k\right)$ | The WB spectral envelope with frequency index $k$ |
| $\tilde{\phi}_{\mathrm{WB}}\left(k\right)$ | The estimation of the WB spectral envelope |
| $k$ | frequency domain bin index |
| $m$ | current frame time index |
| $n$ | time domain sample index |
| $N_a$ | The filter order of the speech source-filter model |
| $N_A$ | Number of cylindrical segments of the acoustic tube |
| $N_{\mathrm{cb}}$ | Number of CB entries |
| $N_f$ | Number of formant frequencies |
| $N_g$ | Number of GMM components |
| $N_s$ | Number of states in the HMM |
| $R_{n_A}$ | Reflection coefficient of acoustic tube with index $n_A$ |
| $s_{\mathrm{NB}}\left(n\right)$ | The received NB speech signal in the time domain |
| $\tilde{s}_{\mathrm{WB}}\left(n\right)$ | The BWE estimation of the WB speech signal in the time domain |
| $S_{\mathrm{NB}}\left(k\right)$ | The received NB speech signal in the frequency domain |
| $S_{n_f,n_A}$ | The sensitivity function value of formant $n_f$ and area coefficient $n_A$ |
| $u\left(n\right)$ | The excitation signal |
| $U_{\mathrm{NB}}\left(k\right)$ | The extracted NB excitation signal in the frequency domain |
| $\mathbf{x}$ | NB speech signal features |

| | |
|---|---|
| $\mathbf{x_1}$ | Frequency-based features |
| $\mathbf{x_2}$ | NB VTAF coefficients |
| $\mathbf{x_3}$ | NB excitation coefficients |

# Abbreviations

| | |
|---|---|
| A/D | Analog to Digital |
| AR | Auto-Regressive |
| ARMA | Auto-Regressive Moving-Average |
| BWE | Bandwidth Extension |
| CB | Codebook |
| D/A | Digital to Analog |
| EM | Expectation-Maximization |
| FFT | Fast Fourier Transform |
| FIR | Finite Impulse Response |
| GMM | Gaussian Mixture Models |
| HB | Highband frequencies - 3400-7000 Hz |
| HMM | Hidden Markov Models |
| HPF | High Pass Filter |
| IFFT | Inverse Fast Fourier Transform |
| IRS | Intermediate Reference System |
| ITU | International Telecommunications Union |
| LB | Lowband frequencies - 50-300 Hz |
| LBG algorithm | Linde, Buzo, Gray algorithm |
| LP | Linear Prediction |
| LPC | Linear Prediction Coefficient |
| LPCC | Linear Prediction Cepstral Coefficient |
| LPF | Low Pass Filter |
| LSD | Log Spectral Distance |
| LSF | Line Spectral Frequencies |

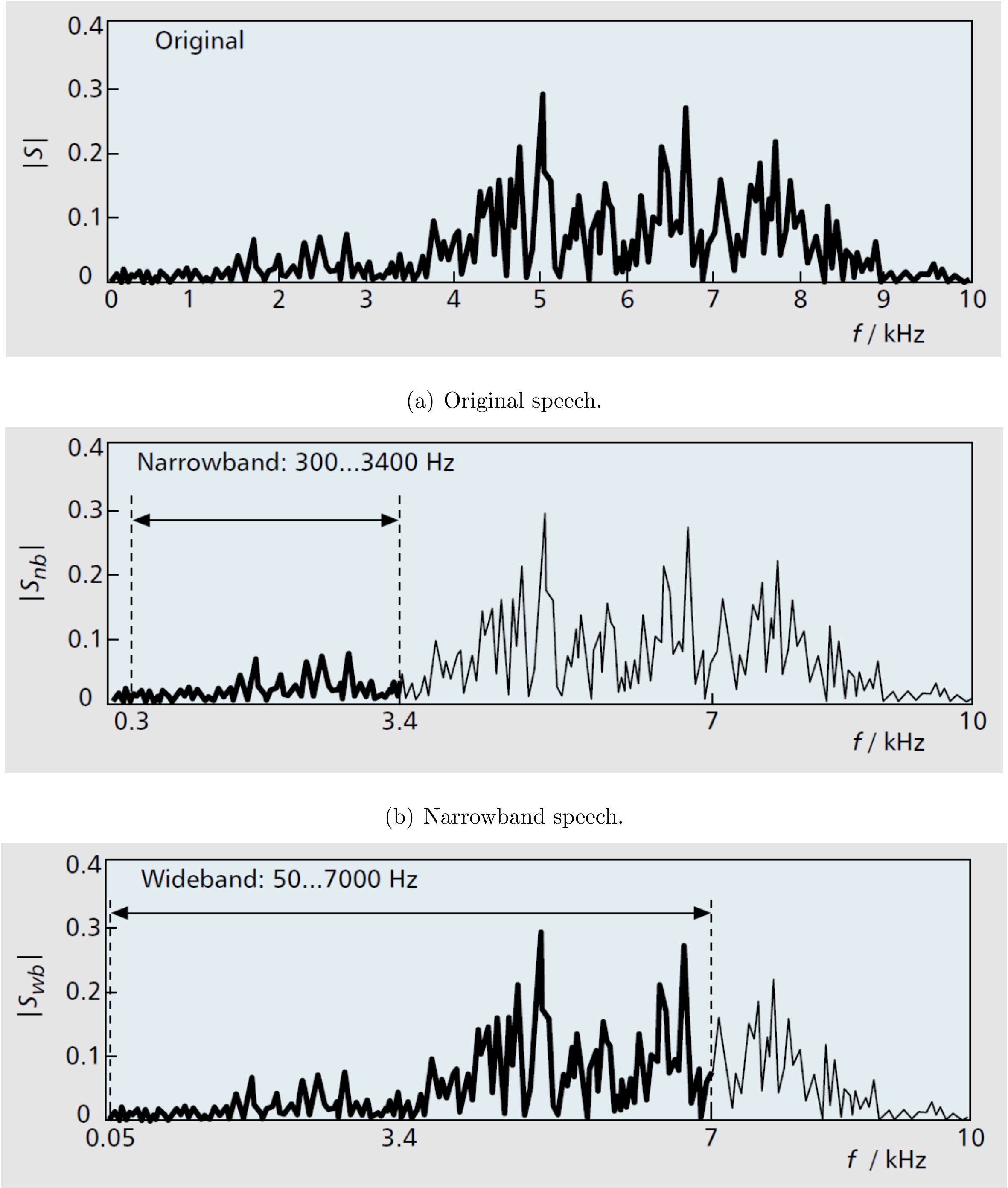
| | |
|---|---|
| MFCC | Mel Frequency Cepstral Coefficients |
| MMSE | Minimum Mean Square Error |
| MOS | Mean Opinion Score |
| MUSHRA | MUltiStimulus test with Hidden Reference and Anchors |
| NB | Narrowband frequencies - 300-3400 Hz |
| PDF | Probability Density Function |
| PSTN | public switched telephone networks |
| SDM | Spectral Distortion Measure |
| SNR | Signal to Noise Ratio |
| SVD | Singular Value Decomposition |
| VQ | Vector Quantization |
| VTAF | Vocal Tract Area Function |
| WB | Wideband frequencies - 50-7000 Hz |

# Chapter 1

# Introduction

## 1.1 Band Limited Speech Signals

Human speech occupies the whole frequency range that is perceptible by the auditory system. An example of an unvoiced speech signal sampled at a sampling frequency of 20 kHz is presented in Fig. 1.1(a). The presented speech sample contains information in the whole frequency range. Limiting the speech bandwidth causes degradation in speech quality, speech naturalness, and speech intelligibility. Fig. 1.2 demonstrates the effect of band limitation on speech intelligibility and speech subjective quality. In Fig. 1.2(a), the intelligibility of meaningless syllables of low-pass and high-pass filtered speech is illustrated. It can be seen that limiting the speech signal to 3.4 kHz yields about 90% intelligibility. This fact makes it sometimes necessary to use the spelling alphabet to communicate words that cannot be understood from the context, for example unknown names. Hence, it increases the telephone network users listening effort. Fig. 1.2(b) compares the speech quality of bandpass-filtered speech with different lower ($f_{c-LB}$) and upper ($f_{c-HB}$) cut-off frequencies. The speech quality is measured in terms of the subjective mean opinion score (MOS), which reflects the subjective rating by human listeners on a scale between one (unacceptable quality) and five (excellent quality). Limiting the speech bandwidth to a narrowband (NB) frequency range from 300 Hz to 3400 Hz yields around 3.2 MOS points. On the other hand, speech band limitation to wideband (WB) frequency range from 50 Hz to 7000 Hz yields around 4.5 MOS points. This 1.3 MOS points gain between NB speech and WB speech yields a significant subjective speech quality improvement.

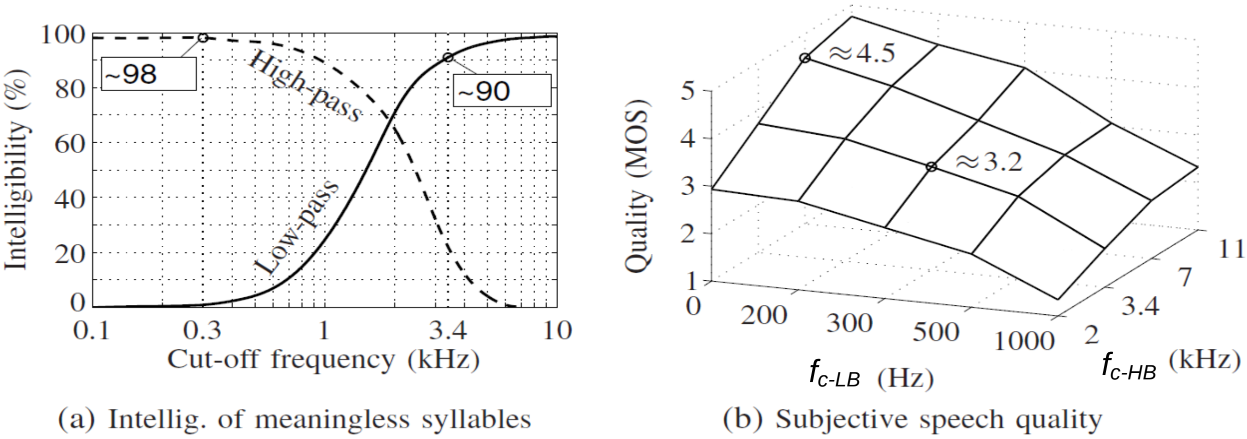


(a) Original speech.

(b) Narrowband speech.

(c) Wideband speech.

Figure 1.1: Short-term spectrum of an unvoiced utterance. $S$: original speech; $S_{\mathrm{NB}}$ narrowband telephone speech; $S_{\mathrm{WB}}$ wideband speech (from [18]).



(a) Intellig. of meaningless syllables

(b) Subjective speech quality

Figure 1.2: Impacts of a bandwidth limitation on speech intelligibility and subjective quality (from [17]).

Current public switched telephone networks (PSTN) limit the bandwidth of the speech signal to 0.3–3.4 kHz. The low frequency cutoff in telephony, 300 Hz, is set to suppress the power line longitudinal interference and other low frequency electrical noises. Typically, there is more than 25 dB attenuation at 50–60 Hz. The upper band boundary, 3400 Hz, is specified to reduce the bandwidth requirements. The filter that specifies the allowed magnitude response of the telephone line which yields the speech band-limitation is described in the ITU-T recommendation P.48 standard [4]. The intermediate reference system (IRS) filter magnitude response is presented in Fig. 1.3. This narrowband (NB) limitation to 0.3–3.4 kHz results in degradation of speech quality and especially in degradation of speech intelligibility and speech naturalness. This degradation also increases the user listening effort. An example of a NB unvoiced speech is presented in Fig. 1.1(b).

One way to achieve high quality speech is by applying a wideband (WB) coding solution. An example of a WB unvoiced speech is presented in Fig. 1.1(c). WB coders, such as the G.722 in digital enhanced cordless telecommunications (DECT) [2] and G.722.2 in mobile phones [3], expand the coded speech bandwidth to 0.05–7 kHz. These WB coders are providing significant improvements in terms of speech intelligibility and naturalness. Unfortunately, this solution requires an expensive network upgrade by operators and phones upgrade by end users. A possible solution for the transition period to WB speech supporting networks, is to artificially extend the NB speech signal to the low-band (LB) frequencies from 50 Hz to 300 Hz and to the high-band (HB) frequencies from 3.4

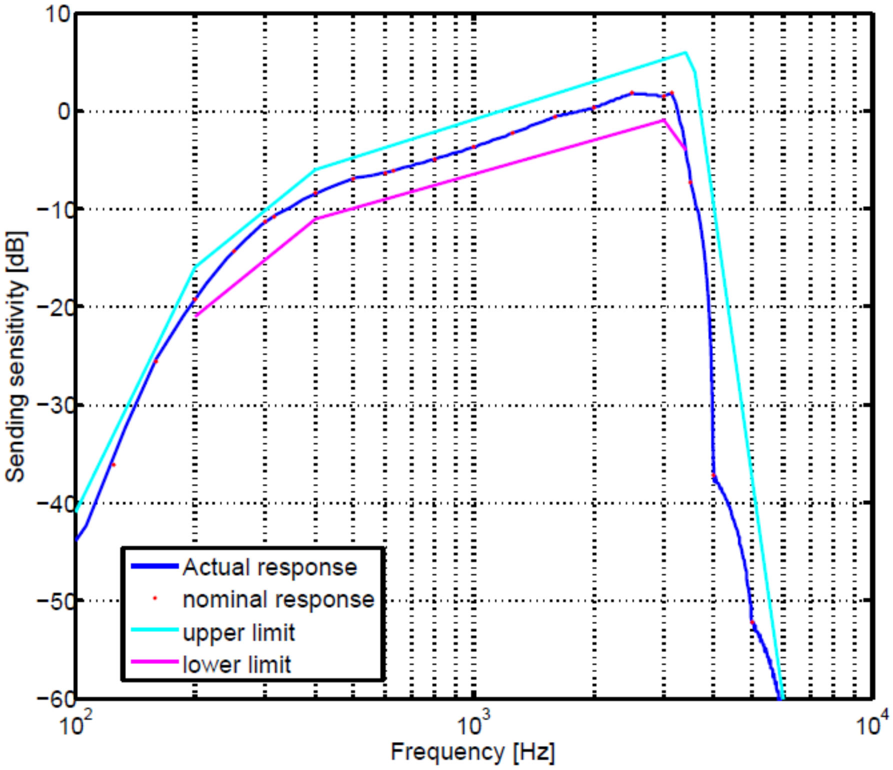


Figure 1.3: The magnitude response of the IRS filter [4].

kHz to 7 kHz [15, 17, 18]. This technique, called bandwidth extension (BWE) of speech, is transparent to the transmitting network, as it is implemented only at the receiving end. Hence, it can be easily and cheaply implemented in the consumers personal phones. A general block diagram showing the location of the BWE unit in the communication network is presented in Fig. 1.4. At the far end there is still a conventional NB telephone with analog-to-digital (A/D) conversion at a sampling rate of $f_s = 8\ kHz$ and a NB coder. At the receiving near-end side the NB speech signal is first decoded using a conventional NB decoder. BWE is then applied to produce a WB signal with a sample rate of $f_s = 16\ kHz$.

A prerequisite for a successful BWE of speech is the sufficient correlation between the speech signal in the NB and the extended regions. Previous work in this field includes model-free BWE algorithms. Yasukawa used spectral folding and a non-linear operator [44, 46] to estimate the high band component and a fixed filter to spectrally shape it. The problem with this method is that it does not achieve good quality for various speech scenarios. Current state-of-the-art BWE of speech algorithms are based on the source-

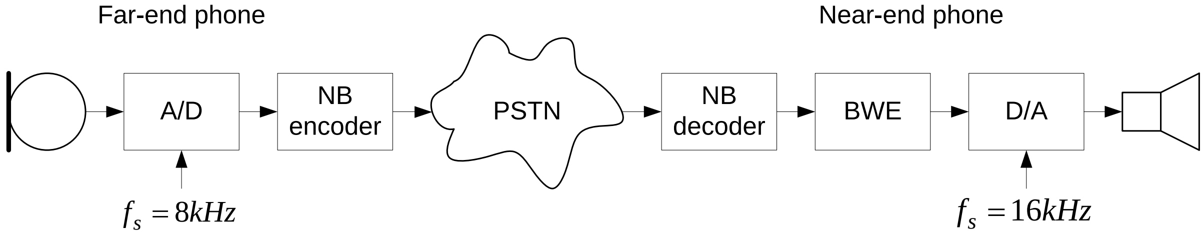


Figure 1.4: Location of BWE algorithm in the communication network.

filter speech model. This model represents the speech signal by a set of parameters. The mutual information between the NB speech parameters and the WB speech parameters can be seen as a measure of how much information is available about the WB speech from the NB speech. High mutual information is therefore desirable. Another requirement is that the representing parameters can be properly separated. The separability is a measure of how well a given feature set can be discriminated, and this is important when doing classification [19]. Model-based BWE algorithms map the NB speech parameters to the WB speech parameters. The mapping techniques are based on vector quantization (VQ) codebook (CB) mapping [10,11,22], linear or piecewise linear mapping [11,13] and statistical methods as neural networks [31,32], Gaussian mixture model (GMM) [28,30,33] and hidden Markov model (HMM) mapping [6,20,38].

## 1.2 Problem Definition and Research Objective

A common solution for BWE algorithm uses the source-filter model of speech production. It assumes that the estimation of missing frequency bands could be divided into two independent tasks of excitation source and vocal tract filter estimation. The current state-of-the-art research in this field addresses two major challenges:

- Estimation of the WB vocal tract filter - this is crucial for high intelligibility speech.
- Estimation of missing bands gain that are matched to the input NB gain - this is crucial for high quality of estimated WB speech without adding any artifacts.

Approaches for tackling the first challenge applied different estimation techniques to map NB speech features to WB speech features, including codebook mapping, linear mapping, and statistical approaches like GMM and HMM. The second challenge is addressed by estimation of missing bands gain directly from the NB speech features or by gain adjustment of estimated WB speech signal to received NB speech signal.

This research considers both challenges: the estimation of WB vocal tract filter and gain adjustment to received NB speech signal. We propose a method for vocal tract filter estimation, based on techniques from the fields of speech recognition and speaker identification. The estimation method includes:

- HMM based estimation of NB speech linguistic content - this allows phoneme dependent estimation of vocal tract filter, which should give better intelligibility of the estimated WB speech signal.

- Codebook search based vocal tract shape estimation of the specific speaker - this allows speaker dependent estimation of vocal tract filter, which should give better intelligibility and improved gain adjustment of the estimated missing bands.

For gain adjustment we propose an iterative technique for tuning the estimated vocal tract shape. This tuning is performed with respect to calculated vocal tract shape from the received NB speech signal. Using the tuned estimated vocal tract shape gives a better gain adjustment of missing bands and reduces artifacts caused from misprediction.

The main research innovation and product are:

- A new method for estimation of the missing bands vocal tract filter. The method comprises an HMM and a codebook search based estimation, and an iterative tuning of the estimated vocal tract filter. This method improves the estimation accuracy, using objective measurements, compared to other estimation methods.

- A modular BWE algorithm from NB speech that results in improved speech quality, as judged by subjective measurements. The algorithm gives better results than other state-of-the-art BWE algorithms using similar estimation techniques, based on speech sound classification and specific speaker characteristics.

## 1.3 Final Report Structure

This thesis is organized as follows:
In Chapter 2, we introduce speech generation models and the speech sound categories.

The parameters that represent the NB and WB speech and the parameter extraction techniques are presented. The suitability of the NB parameters for intelligent estimation of the missing bands is presented.

In Chapter 3, we present some existing methods of BWE of NB speech. We introduce the general framework of all BWE algorithms and the major problems which are still tackled in this field.

In Chapter 4, we describe the proposed BWE algorithm. We present the general algorithm structure and we describe in detail the algorithm stages including:

- Preprocessing and feature extraction stage.
- WB spectral envelope estimation stage.
- WB excitation generation stage.
- WB speech synthesis stage.

In Chapter 5, we evaluate the proposed BWE algorithm performance. We examine both objective and subjective measurements to evaluate the algorithm strength and needed future work.

In Chapter 6, we draw our conclusions and discuss some possible future work.

# Chapter 2

# Speech Modeling and Analysis

In order to make intelligent estimation of the missing bands, some assumptions and properties of the speech signal must be considered. The first assumption to consider is the speech production model. The knowledge that the parameters that represent the target WB and the received NB speech signals stem from the same assumed model, allows better estimation of the WB speech parameters. Another property to consider is the various types of sounds that make up human speech. This knowledge provides information on the relative importance of different speech frequency bands.

The speech model and speech properties are characterized by different parameters. Each NB parameter conveys different amount of information on the WB speech signal. The chosen parameters that represent the NB speech affect the quality of the estimated BWE speech signal and the complexity of the BWE algorithm.

In this chapter we will present the speech generation models and the speech sound categories. The parameters that represent the NB and WB speech and the parameter extraction techniques are presented. The suitability of the NB parameters for intelligent estimation of the missing bands is discussed.

## 2.1 Speech Generation Models

Speech is an acoustic waveform that conveys information from a speaker to a listener. The human vocal system is described in Fig. 2.1. Air exhaled from the lungs passes through the vocal chords at the larynx and then through the vocal tract which consists of

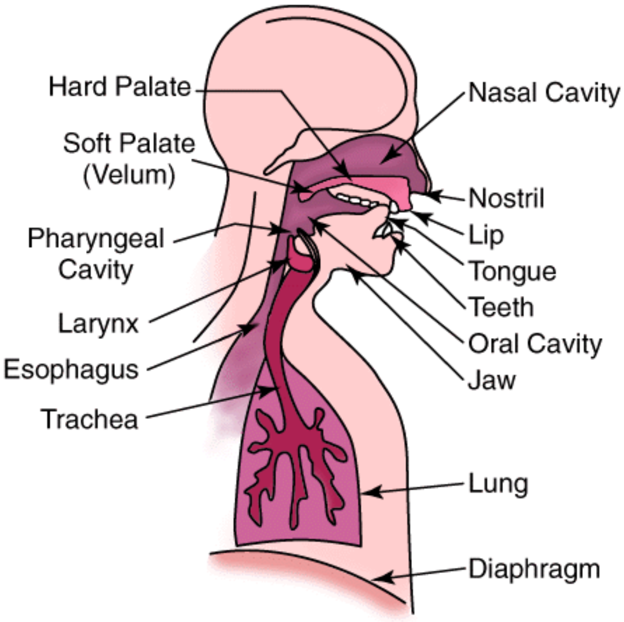


Figure 2.1: Diagram of the speech production system (from [16]).

the pharyngeal, nasal and oral cavities. This air flow is shaped into certain characteristics to produce the speech sounds. The speech sounds can be categorized into two major categories according to the mode of air excitation of the vocal tract. *Voiced* sounds are produced by forcing air through the larynx, with the tension of the vocal cords adjusted so that they vibrate in a relaxed oscillation. The produced quasi-periodic pulses of air excite the vocal tract to produce the voiced speech sounds. The period of the excitation for voiced sounds, known as the pitch period, determine the pitch frequency. *Unvoiced* sounds are produced by turbulence, as air is forced through a constriction at some point in the vocal tract. Voiced sounds are characterized by high energy which is mostly concentrated in the NB frequency band whereas unvoiced sounds are characterized by low energy which is mostly concentrated in the HB frequency band.

Speech sounds can be further classified into phoneme classes. A phoneme is a basic speech sound unit of a particular language capable of conveying a distinction in meaning. The different phoneme sounds are characterized by the air flow excitation and the vocal tract shape [35]. *Vowel* phonemes are characterized by quasi-periodic air pulses that excite the vocal tract. The shape of the cavities that comprise the vocal tract, known as the vocal tract area function (VTAF), determines the resonance frequencies, also known as formants, which are emphasized in the specific vowel phoneme. An example of a vowel is /A/ as in "h*a*t". *Diphthong* phonemes are characterized by a time varying VTAF which varies between two vowels configurations. An example of a diphthong is /oU/ as in "b*oa*t". *Semivowel* phonemes are characterized by a gliding transition in VTAF between adjacent phonemes. An example of a semivowel is /L/ as in "*l*ight". *Nasal* phonemes are characterized by glottal excitation while the vocal tract totally constricted at some point along the oral cavity. An example of a nasal is /N/ as in "*n*ight". *Unvoiced fricative* phonemes are characterized by a steady air flow excitation which becomes turbulent or noise like excitation in the region of a constriction in the vocal tract. An example of an unvoiced fricative is /S/ as in "*s*everal". *Voiced fricative* phonemes are characterized by semi-periodic excitation formed by vibrating vocal cords that become turbulent or noise like excitation in the region of a constriction in the vocal tract. An example of a voiced fricative is /Z/ as in "*z*ero". *Voiced stop* phonemes are characterized by a sudden release of air pressure which is building up behind a total constriction somewhere in the oral cavity. The vocal cords are able to vibrate although the vocal tract is closed at some point. An example of a voiced stop is /B/ as in "*b*ear". *Unvoiced stop* phonemes are similar to the voiced stops with the one exception that the vocal chords do not vibrate. An example of an unvoiced stop is /T/ as in "ea*t*". *Affricate* phonemes are consisting of a stop consonant followed by a fricative, and hence are characterized by an air flow exciting a constraint at the vocal tract. An example of an affricates is /J/ as in "*j*oin".

An example of a voiced vowel /A/ and an unvoiced fricative /S/ in the time and frequency domains is presented in Fig. 2.2. It can be seen that the voiced phoneme have semi-periodic form and high energy concentrated in the NB frequencies below 4 kHz. The unvoiced phoneme is characterized with lower energy which is mostly concentrated in the HB frequencies above 4 kHz.

Speech production is simply modeled by the source-filter model as air passing thorough the vocal tract [9, 35]. This model considers the speech signal as being produced by a spectrally flat excitation source that passes through an auto-regressive (AR) filter. The excitation source corresponds to air flowing from the human lungs through the vocal cords. The AR filter corresponds to the human vocal tract and is also known as the spectral envelope filter. The source-filter model is described in Fig. 2.3. For voiced

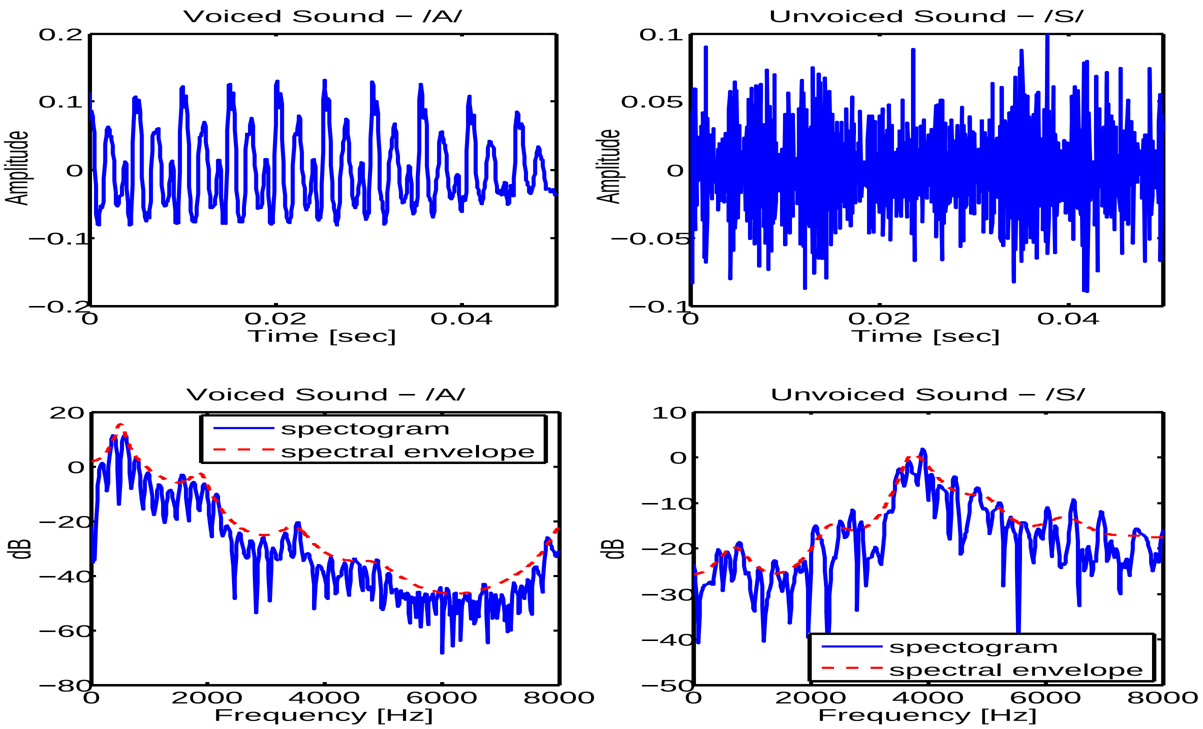


Figure 2.2: Acoustic waveform and spectrum of voiced and unvoiced phonemes.

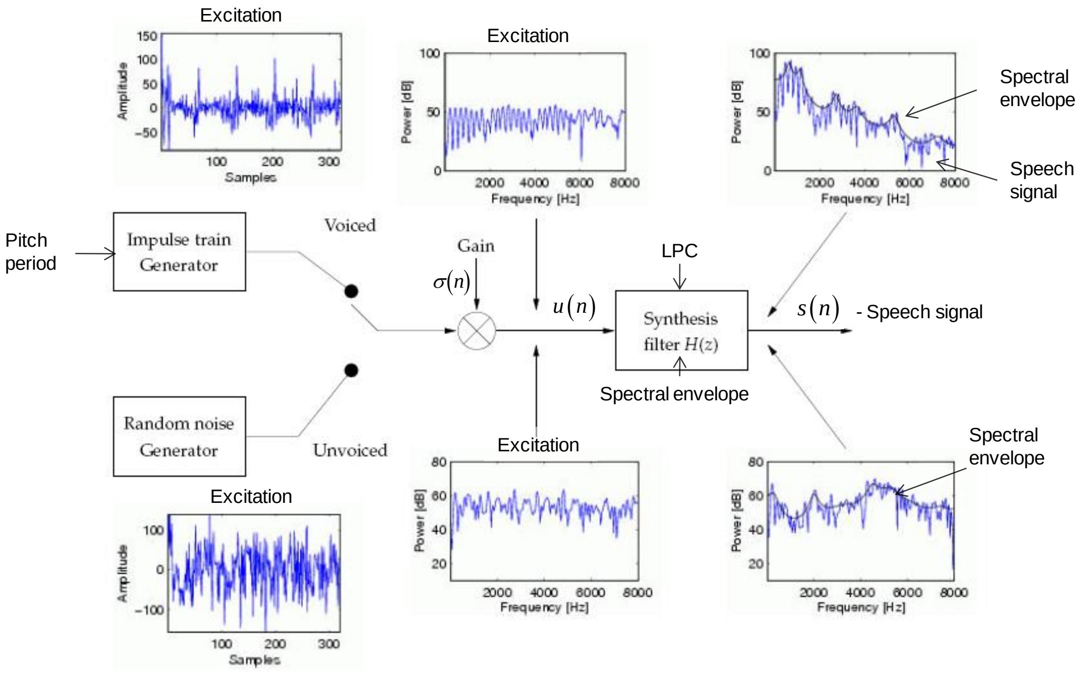


Figure 2.3: Block diagram of the source-filter model.

speech sounds, the excitation is modeled by an impulse train with given pitch frequency. For unvoiced speech sounds, the excitation is modeled by a random noise generator. The excitation signal is multiplied by the excitation gain. Using linear prediction (LP) analysis to represent the source-filter model, the speech signal can be expressed as a combination of weighted previous speech samples and the excitation signal:

$$s\left(n\right)=\sum_{n_a=1}^{N_a}a_{n_a}s\left(n-n_a\right)+\sigma\left(n\right)u\left(n\right), \tag{2.1}$$

where $s\left(n\right)$ is the speech signal with time index $n$, $a_{n_a}$ is the linear prediction coefficients (LPC) based AR filter that represents the vocal tract and $N_a$ is the filter order. $u\left(n\right)$ is the excitation signal and $\sigma\left(n\right)$ is the excitation gain.

The physical action of the source-filter model can be investigated using the acoustic theory, which represents the motion of air in the vocal system by partial differential equations. A simple way to solve these equations is to model the vocal tract as a concatenation of acoustic tubes of non-uniform, time varying, cross section. Atal and Wakita showed in [5, 42, 43] that the LP based AR filter that represent the vocal tract is equivalent to a non-uniform acoustic tube. An acoustic tube length $L$ consists of $N_A$ cylindrical segments as illustrated in Fig. 2.4. All segments have the same length $l = L/N_A$ but different cross sectional areas $A_{n_A},\ n_A = 1, \ldots, N_A$. Atal and Wakita showed that the number of cylindrical segments, $N_A$, is equal to the order of the AR filter, $N_a$. The model order $N_A$ depends on the sampling rate $f_s$. It is derived by solving the sound wave propagation equation at the segments boundaries using a continuity constraint. The model order is chosen to fulfill the following constraint:

$$N_A = f_s\frac{2L}{c}, \tag{2.2}$$

where $c$ is the speed of sound in the vocal tract. Good choice for the VTAF order is $N_A = 8$ which corresponds to $f_s = 8000Hz$ , $c = 34000cm/sec$ and $L = 17cm$ which is the typical length of the vocal tract for adults. The VTAF is calculated by matching the pressure and the volume velocity at the junction between adjacent sections [35]. This yields the following recursive formula for the area sections:

$$A_{n_A} = \frac{1+R_{n_A}}{1-R_{n_A}}A_{n_A+1}, \tag{2.3}$$

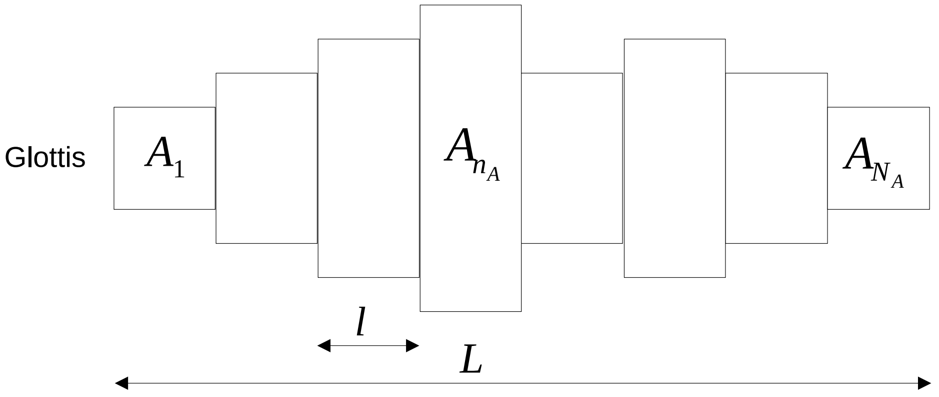


Figure 2.4: Vocal tract model: acoustic tube without loss consisting of cylindrical segments of equal length.

where $R_{n_A}$ is the reflection coefficient. The initial area value $A_{N_A+1}$ is arbitrary set to 1. The reflection coefficients can be derived as a byproduct of the Levinson-Durbin algorithm which aims to calculate the LP coefficients [35].

As was shown in [42,43], the VTAF approximates the speaker's physical speech production shape. The VTAF indicates the locations of the formant frequencies that contribute to both phonetic and speaker-specific characteristics. One application using the VTAF is speech modification of formants location by VTAF perturbation [40]. The modification of formants to a specific pattern is produced by iteratively adjusting the shape of a given VTAF using the sensitivity function. The sensitivity function relates small changes in VTAF to changes in formant frequencies. The motivation for this technique is in the fact that a subtle VTAF perturbation causes the required modification of formants.

We denote the VTAF values by $A_{n_A}$, $n_A = 1, \ldots, N_A$, where $N_A$ is the number of area coefficients. The formant frequencies are denoted by $f_{n_f}$, $n_f = 1, \ldots, N_f$, where $N_f$ is the number of formant frequencies. The sensitivity function $S_{n_f,n_A}$ satisfies the following relationship:

$$\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A=1}^{N_A} S_{n_f,n_A} \frac{\Delta A_{n_A}}{A_{n_A}}, \tag{2.4}$$

where $\Delta f_{n_f}$ is the difference between the desired formant frequency and the current formant frequency, and $\Delta A_{n_A}$ is the perturbation size of the ${n_A}^{th}$ segment area. This equation says that for specific formant $n_f$ and specific segment $n_A$, if the sensitivity function is positive valued and the area perturbation is also positive (area is increased), the change

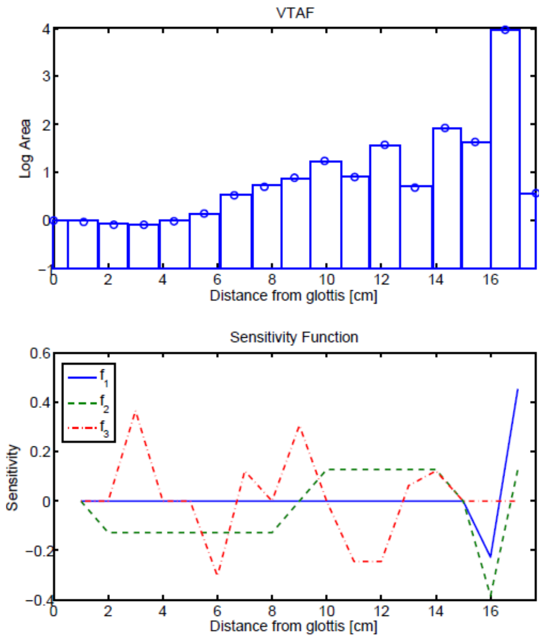


Figure 2.5: An example of calculated VTAF and its formants-area sensitivity function.

in formant frequency will be upward (positive). If the area change is negative (area is decreased), the formant frequency will decrease. When the sensitivity function is negative, the opposite effect occurs for positive or negative area perturbations. An example of a VTAF (in log scale) and its sensitivity function is presented in Fig. 2.5.

Another possible use of the LP based VTAF in speech processing can be found in the fields of speech analysis, speech recognition, speech coding, and speech enhancement [7, 23, 24, 26, 37]. The model limitations are as follow:

- The nasal tract is not included. Hence, nasal vowels are modeled inaccurately.
- The production of unvoiced plosives and fricatives are not included.
- The losses at the wall of the vocal tract and other losses are not modeled. The model assumes sound waves progressing in the vocal tract are plane waves. These might cause inaccurate VTAF estimation.
- The effects of glottal waveform shape and radiation at the lips are not part of the model. Hence, pre-emphasis of the speech signal should be performed to compensate for those effects.

Despite these limitation, this model of speech production allows good geometric shape mapping of VTAF for most of speech utterances [23].

## 2.2 Speech Parameters Calculation

The speech signal can be modeled by the excitation source and the spectral envelope filter as depicted in the source-filter model. LP analysis represent the speech, as seen in equation (2.1), from weighted past speech samples. From this equation, the excitation is extracted by:

$$u\left(n\right)=s\left(n\right)-\sum_{n_a=1}^{N_a}a_{n_a}s\left(n-n_a\right). \tag{2.5}$$

The following parameters characterize the excitation signal:

**Pitch Frequency**

The pitch frequency denotes the harmonics of the source pulse train. The pitch frequency can be estimated by various algorithms described in [15, 35].

**Voicing Degree**

The voicing degree weights the pulse train and noise generated excitation signals to sum to 1. This excitation mix creates a new more realistic excitation signal.

**Excitation Gain**

The gain of the excitation signal.

The spectral envelope has several known parametric representations. Most of the representing parameters can be derived from the **LPC** which construct the AR filter. LPC can be calculated by the autocorrelation method using the Levinson-Durbin algorithm and by the covariance method [35]. These methods use different definitions of the segments of the signal to be analyzed and hence yield LPC with somewhat different properties. We presented in the last section the connection between the LPC and the **VTAF** parameters through the **reflection coefficients**. Other representation includes the **line spectral frequencies (LSF)** and the **linear prediction cepstral coefficients (LPCC)** [15].

Another set of coefficients that are very popular in the field of speech recognition are the **mel-frequency cepstral coefficients (MFCC)**. The MFCC represent the short term power spectrum of a sound, based on a linear cosine transform of a log power

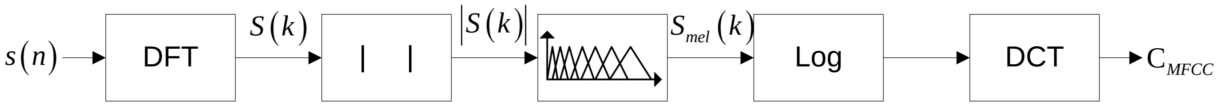


Figure 2.6: Block diagram of the MFCC calculation.

spectrum on a non-linear mel scale of frequency. MFCC are derived as follows (see block diagram at Fig. 2.6):

- Calculate the Fourier transform of a windowed excerpt of a signal.
- Map the magnitudes of the transform values obtained above onto the mel scale, using triangular overlapping windows.
- Take the log of the magnitudes at each of the mel frequencies.
- perform the discrete cosine transform of the obtained log magnitudes.
- The MFCC are the resulting real coefficients.

The following scalar speech features allow to categorize the analyzed speech frame to different sound types:

**Spectral Centroid**

The spectral centroid is a measure that indicates where most of the power of a speech frame is spectrally located. It is generally high for unvoiced sounds [15]. The normalized spectral centroid is defined as:

$$\mathbf{x_{SC}} = \frac{\sum\limits_{k=0}^{N_{FFT}/2} k\left|S\left(k\right)\right|}{\left(\frac{N_{FFT}}{2}+1\right)\sum\limits_{k=0}^{N_{FFT}/2}\left|S\left(k\right)\right|}. \tag{2.6}$$

where $N_{FFT}$ is the size of the fast Fourier transform (FFT). The $\mathbf{x_{SC}}$ range is between 0 to 1.

**Spectral Flatness Measure**

The spectral flatness measure indicates the tonality of the speech signal [31]. The meaning of tonal in this context is in the sense of the peakedness or resonant structure of the power spectrum, as opposed to the flat spectrum of a white noise. A high spectral flatness value indicates that the spectrum has a similar amount of power in all spectral bands and the graph of the spectrum would appear relatively flat and smooth. This property is mostly common with unvoiced sounds. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands and the spectrum would appear "spiky". This property is mostly common with voiced sounds. The spectral flatness measure range between 0 to 1 and is defined as:

$$\mathbf{x_{SF}} = \frac{\left(\prod_{k=0}^{N_{FFT}-1} k\left|S\left(k\right)\right|^{2}\right)^{\frac{1}{N_{FFT}}}}{\frac{1}{N_{FFT}}\sum_{k=0}^{N_{FFT}-1}\left|S\left(k\right)\right|^{2}}. \tag{2.7}$$

### Spectral Slope

The spectral slope which is useful for discriminating voiced frames from fricatives and plosives [32]. It is calculated by dividing the mean FFT magnitude in a high-band frequency from 2800-3400 Hz with a low-band frequency from 400-1000 Hz:

$$\mathbf{x_{SS}} = \frac{\sum_{k=\mathbf{K_{high}}}\left|S\left(k\right)\right|}{\sum_{k=\mathbf{K_{low}}}\left|S\left(k\right)\right|}, \tag{2.8}$$

where $\mathbf{K_{high}}$ and $\mathbf{K_{low}}$ represent the frequency bins of the frequency bands described above.

### Short Term Energy

The short term energy for frame $m$ with $N$ samples is defined as:

$$E\left(m\right) = \sum_{n=0}^{N-1} s^{2}\left(n\right). \tag{2.9}$$

### Long Term Energy

The long term energy for frame $m$ is recursively calculated as:

$$\tilde{E}\left(m\right) = \alpha\tilde{E}\left(m-1\right) + \left(1-\alpha\right)E\left(m\right). \tag{2.10}$$

### Background Noise Energy

The background noise energy can be simply estimated by:

$$E_{min}\left(m\right) = \min_{\mu=0}^{N_{min}} E\left(m-\mu\right), \tag{2.11}$$

where $N_{min}$ is the number of searched frames.

**Normalized Relative Frame Energy**

The normalized relative frame energy is calculated by normalizing the short term energy with the long term energy to receive independent measure to different speakers, or different recording, or transmission equipment [15]. The normalized energy is defined as:

$$\mathbf{x_{NRFE}} = \frac{10log_{10}E\left(m\right) - 10log_{10}E_{min}\left(m\right)}{10log_{10}\tilde{E}\left(m\right) - 10log_{10}E_{min}\left(m\right)}. \tag{2.12}$$

Fig. 2.7 shows some of the above scalar features for a specific utterance. The upper graph presents the WB and NB signals in the time domain. The second graph presents the NB signal spectrogram. In the spectrogram it is possible to see voiced speech parts characterize by high energy (dark red) in the low frequencies and unvoiced speech parts characterize by low energy (yellow) in the low frequencies. It is seen that the spectral centroid $\mathbf{x_{SF}}$ increases during unvoiced and plosive sounds. The spectral flatness $\mathbf{x_{SF}}$ shows similar trend to the spectral centroid although it seems to give more information on a frame by frame basis changes. The spectral slope $\mathbf{x_{SS}}$ shows the same trend as the spectral centroid, but it is not entirely equal, therefore it should bring additional valuable information. The normalized relative frame energy $\mathbf{x_{NRFE}}$ seems to be a good indication of voice activity and it also distinguishes between the voiced/unvoiced sections.

The potential quality of each parameter for the missing bands estimation was investigated in [19, 29, 30]. The quality of each feature is quantified in terms of the statistical measures of mutual information and separability (Appendix C. Shannons mutual information, $I\left(\mathbf{x}:\mathbf{y}\right)$, between the feature set $\mathbf{x}$ and the estimated quantity $\mathbf{y}$ can be regarded as an indication of the feasibility of the estimation task. From the field of pattern recognition the separability, $\zeta\left(\mathbf{x}\right)$, is known as a measure for the quality of a particular feature set for a classification and estimation problem [12].

Table 2.1 summarizes some of the results in [19]. The presented estimate results of mutual information $I\left(\mathbf{x}:\mathbf{y}\right)$ and separability $\zeta\left(\mathbf{x}\right)$ for BWE of the high frequency band (3.4-8 kHz) from telephone speech (0.3-3.4 kHz) allow making an intelligent selection of parameters for the BWE algorithm. To achieve the best results with the BWE algorithm, it motivates using the MFCC feature vector, which yield the highest separability and one

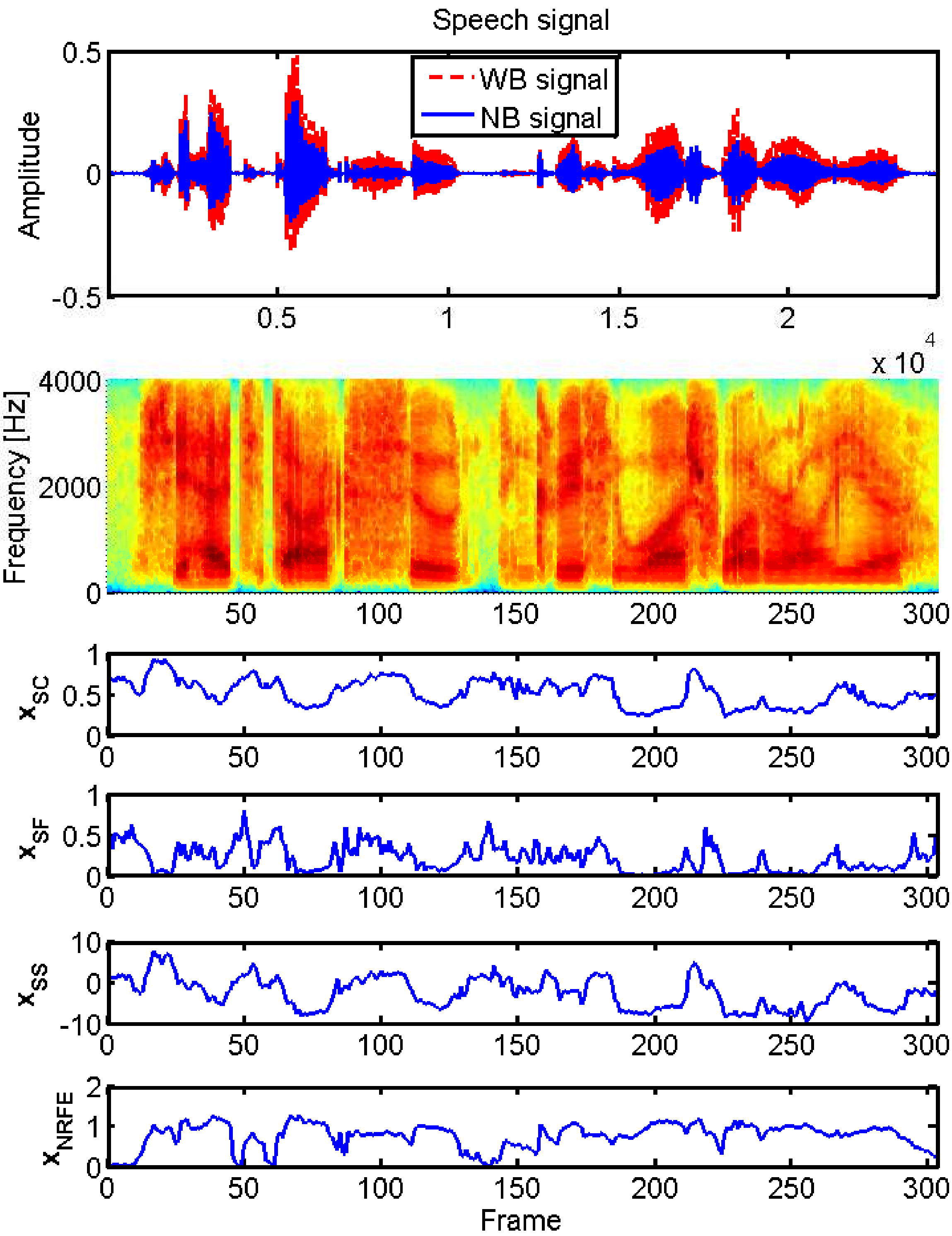


Figure 2.7: Scalar features of a NB speech signal.

Table 2.1: Mutual information and separability for NB speech features (from [19]).

| feature vector $x$ | dim $x$ | $I\left(x : y\right)$<br>[bit/frame] | $\zeta\left(x\right)$<br>(16 classes) |
|---|---|---|---|
| LPC | 10 | 2.3054 | 1.5295 |
| LSF | 10 | 2.3597 | 1.5596 |
| LPCC | 10 | 2.2401 | 1.4282 |
| MFCC | 10 | 2.3325 | 2.2659 |
| frame energy | 1 | 0.9285 | 1.0756 |
| pitch period | 1 | 0.2451 | 0.0530 |
| spectral centroid | 1 | 0.7913 | 1.0179 |
| spectral flatness | 1 | 0.4387 | 0.3538 |

of the highest mutual information between all possible features. Additional scalar feature that yield high mutual information is the frame energy.

## 2.3 Summary

We have presented the speech generation models. The physical speech production process can be modeled by an excitation source signal passing through the vocal tract filter. The concatenated tube model represents the physical shape of the speaker's vocal tract. The speech can be classified into various phonemes that are characterized by the properties of the excitation and the vocal tract shape.

Various parameters represent the excitation and the spectral envelope signals. From those features we can analyze the speech characteristics and the amount of information in the missing HB frequencies. It is seen that the MFCC feature vector and the energy scalar feature, extracted from the NB signal, yields high mutual information to the HB signal, and high separability. These features could allow improved BWE algorithm results.

In the next chapter we will present several known BWE algorithms and examine the different speech models and speech parameters used.

# Chapter 3

# Methods of Bandwidth Extension of Speech

The aim of BWE algorithms is to improve NB speech quality, naturalness and intelligibility by estimating the speech signal at the missing low and high frequency bands. In this chapter methods of BWE of speech are reviewed.

Different approaches for BWE of speech exist and many give promising results. The approaches differ in the algorithm framework, which is affected by the choice of the speech model, whether the processing is done in the time domain or in the frequency domain, and the method of analysis and synthesis. They also differ in the method of estimation of the missing frequency bands.

The major problem of most existing BWE methods is the suitable estimation of the HB spectral envelope, which affect the quality and intelligibility of the estimated speech signal. Another major problem is the estimation of the extended bands signals gains, which affect the amount of artifacts in the synthesized signal.

## 3.1 General Framework

Generally, a BWE algorithm can be divided into four stages (See Fig. 3.1):

- Spectral component generation of the missing frequency bands.
- Spectral shaping of the missing frequency bands generated components.

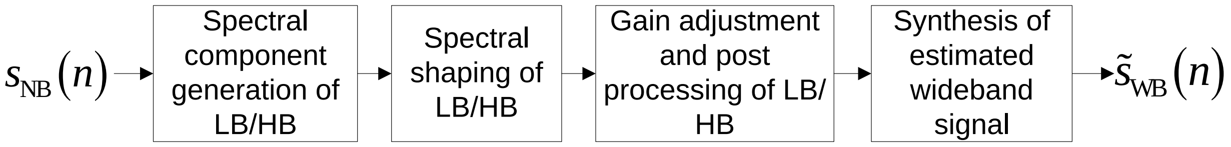

$s_{NB}(n)$
Spectral component generation of LB/HB
Spectral shaping of LB/HB
Gain adjustment and post processing of LB/HB
Synthesis of estimated wideband signal
$\tilde{s}_{WB}(n)$

Figure 3.1: Block diagram of general BWE algorithm.

- Gain adjustment of estimated missing bands signals to match the NB signal gain.
- Synthesis of WB signal from existing NB and estimated signals in missing frequency bands.

All BWE algorithms include these stages, with the major difference among them is the chosen speech model that the algorithm uses. The most basic BWE method is model-free. I.e., it is not using any speech model. This is known as the non-model based BWE of speech. Yasukawa used spectral folding and a non-linear operator [44, 46] to estimate the high band component and a fixed filter to spectrally shape it. The non-model based algorithm in [32] uses a genetic algorithm search technique and a cubic-spline interpolation to estimate the shaping filter in the frequency domain. This is performed by estimating several HB spectral shaping filter magnitude values using a genetic algorithm, and applying cubic-spline interpolation to these values to construct the full shaping filter. This algorithm achieves a good quality, low complexity, BWE algorithm. Although the non-model methods are of low complexity and have low dependency on speech parameters and channel conditions, they do not provide good bandwidth extension quality for various speech scenarios.

Most recent BWE algorithms use the source-filter model of speech production. This model considers the speech signal as being produced by a spectrally flat excitation source that passes through an auto-regressive (AR) filter [35]. This model suggests separation of the BWE algorithm into two independent tasks of excitation and spectral envelope estimation [15]. A general model-based BWE block diagram is described in Fig. 3.2. The NB speech signal is analyzed for the NB excitation and spectral envelope signals. Estimation of the excitation signal and the spectral envelope of the missing bands is performed next. Gain adjustment of the estimated signals to match the received NB signal follows. The final BWE stage is the synthesis of the estimated BWE speech signal by

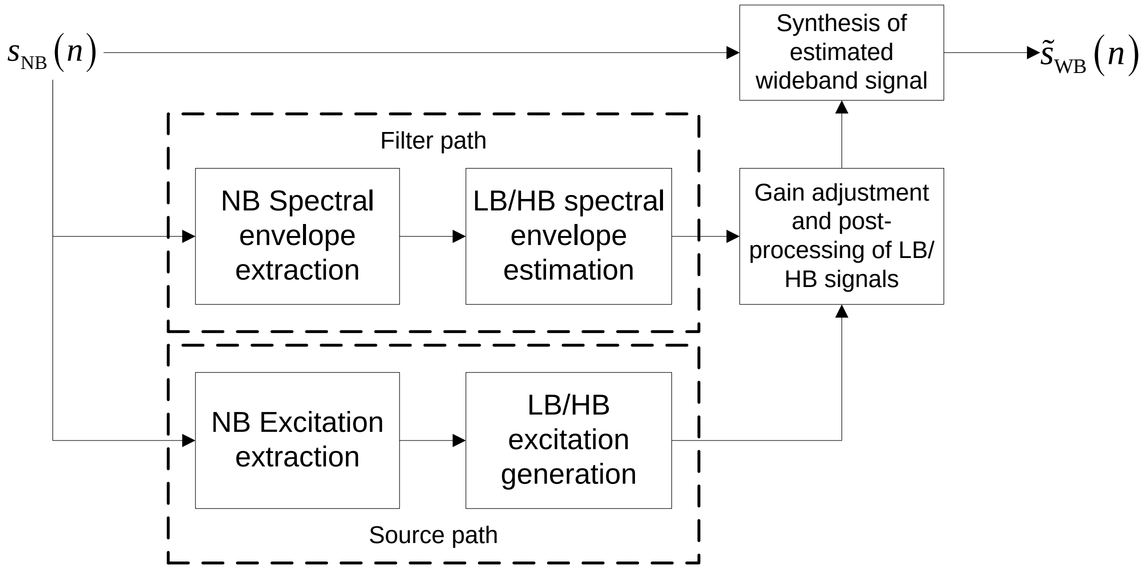

$s_{\mathrm{NB}}(n)$
Synthesis of estimated wideband signal
$\tilde{s}_{\mathrm{WB}}(n)$
Filter path
NB Spectral envelope extraction
LB/HB spectral envelope estimation
Gain adjustment and post-processing of LB/HB signals
NB Excitation extraction
LB/HB excitation generation
Source path

Figure 3.2: Block diagram of model-based BWE algorithm.

concatenation of the missing bands signals with the NB signal. The excitation and spectral envelope estimation tasks correspond to the first two stages of the general BWE scheme. The motivation for using this approach is that it allows the estimation of excitation and spectral envelope without any dependency. Another motivation is that the NB excitation and spectral envelope parameters have some correlation to those of the WB signals as was described in the previous chapter.

The model-based approach can be further divided into algorithms that use an offline training procedure as an a-priori information for estimation and those that do not. In most BWE algorithms training is used to map NB to WB features. These features mostly represent the speech frame spectral envelope, energy ratio and other speech classification features. The mapping techniques are based on vector quantization (VQ) codebook (CB) mapping [10,11,22], linear or piecewise linear mapping [11,13] and statistical methods as GMM and HMM mapping [20,28,33].

## 3.2 Excitation Generation

The excitation generation of the LB and HB is one of the two major sub-systems in any model-based BWE algorithm. This sub-system calculates the NB excitation signal, estimates the missing LB and HB excitation, and generates the WB excitation which serves as the input to the final estimated spectral envelope filter. There are many methods to derive the missing excitation which can be divided into three categories:

- *Spectral shifting*, which includes spectral folding, modulation and spectral copy. These techniques can be applied in the time and frequency domain. They use the NB excitation as input.
- Applying a *non-linear operator*, consisting of a rectifier or a quadratic function. These techniques maintain the harmonic structure of the excitation and are applied in the time domain. They also use the NB excitation as input.
- *Function generation*, using a sinusoidal tone generator, mostly for voiced sounds and for LB generation, and a white noise generator, mostly for unvoiced sounds and for HB generation. These techniques use analyzed parameters from the NB signal.

Spectral folding [10, 32] is probably the simplest way to generate the HB excitation. It is performed by upsampling the NB excitation signal without passing it through an anti-aliasing LPF.

$$\tilde{u}_{\mathrm{WB}}(n) = \begin{cases} u_{\mathrm{NB}}(n) & , \ \text{n odd} \\ 0 & , \ \text{otherwise} \end{cases} , \tag{3.1}$$

where $u(n)$ represents the excitation signal. This operation results in folding of the power spectrum. Its drawbacks are the spectral gap from 3.4 kHz to 4.6 kHz and the failure to maintain harmonic structure for voiced sounds. Its advantages are in its simplicity and in the fact that it maintains the NB excitation unchanged. NB excitation modulation in the time domain [20, 22], and spectral copy in the frequency domain [13] shift the excitation content of the NB signal to the HB. This excitation shift allows solving the spectral gap of the spectral folding method. The advantage of spectral copy over modulation is in the maintenance of the NB excitation unchanged. Excitation modulation with specific modulation frequency $\Omega$ is calculated as:

$$\tilde{u}_{\mathrm{HB}}(n) = \mathrm{HPF}\left\{u_{\mathrm{NB}}(n) \cdot 2\cos(\Omega n)\right\} \quad , \tag{3.2}$$

where HPF{} is an high pass filter of the modulated excitation. Pitch-synchronize modulation and peak picking based spectral copy also allow maintaining the harmonic structure of the excitation in the HB. The disadvantage of the pitch-synchronize modulation is in its high complexity caused by the pitch calculation.

Non-linear operators also maintain the harmonic structure of the NB excitation signal. Full-wave rectification of the NB excitation extends its bandwidth to the entire WB [36]:

$$\tilde{u}_{\mathrm{WB}}(n) = |u_{\mathrm{NB}}(n)| . \tag{3.3}$$

Using LP and HP filters allow extracting the LB and HB excitation appropriately. Quadratic function and LPF are used in [22] for LB excitation extension:

$$\tilde{u}_{\mathrm{LB}}(n) = \mathrm{LPF}\left\{u_{\mathrm{NB}}^{2}(n)\right\} . \tag{3.4}$$

The drawbacks in these methods are in the fact that they color the generated excitation signal. This coloring of the excitation requires further processing by a whitening filter. They also change the original NB excitation.

Excitation generation using generating functions is described in Fig. 3.3. The NB excitation is analyzed for the pitch frequency, voiced/unvoiced classification and signal gain. During unvoiced speech frames, only the HB part of the excitation is generated as the LB part contains minimal information. The white noise generator generates an excitation signal which is then processed using a HPF. During voiced speech frames, the sine generator generates sinusoidal tones at the pitch harmonics in the LB and HB frequencies range. The calculated NB excitation gain is used to adjust new generated excitation components to match the received NB excitation gain. This excitation generation method implements directly the source generation from the source-filter speech model. Its drawback is the dependence on the quality of the excitation parameters estimation.

Generally, most research made in the field of excitation extension for BWE algorithms showed that the chosen method for excitation generation has little effect on the overall BWE algorithm quality. Hence, a simple excitation generation method that has as many advantages as possible should be a reasonable choice.

## 3.3 Spectral Envelope and Gain Estimation

The estimation of the HB spectral envelope and its gain is the most crucial stage for a high quality BWE algorithm. The HB extension of the spectral envelope aims to

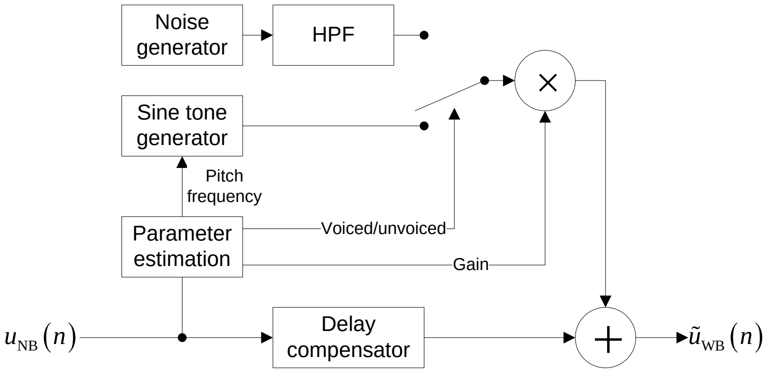


Figure 3.3: Block diagram of excitation generation using generating functions (from [17]).

enhance speech quality, as well as intelligibility. The HB spectral envelope gain may affect the level of artifacts, interpreted as quality degradation. Hence, most recent BWE algorithms use different techniques to map NB speech features to HB features that represent the HB spectral envelope and gain. These techniques include vector quantization (VQ) based codebook (CB) mapping [11,22], linear mapping [13,36], neural networks [32] and statistical methods by Gaussian mixture models (GMM) and hidden Markov models (HMM) [20,30]. These techniques still face problems with spectral envelope estimation of some speech sound classes, especially unvoiced sounds. They also show quality variations for different speakers and some hissing and whistling artifacts due to gain overestimation and discontinuities in the time evolution of the estimated spectral envelope.

## 3.3.1 Highband Spectral Envelope Estimation

### Estimation Based on Speech Sound Classification

One method to improve HB spectral envelope estimation is to incorporate speech-sound class information in the estimation process. This information is especially crucial for better estimation of unvoiced sounds, which are characterized by low NB energy while having high HB energy.

Voiced and unvoiced sounds classification is used in [11] for codebook mapping. CB mapping relies on a one-to-one mapping between two codebooks of NB and HB spectral envelopes to predict the HB envelope. The NB and HB codebooks are trained jointly using the Linde, Buzo and Gray (LBG) algorithm [15], where the spectral envelope parameters of the NB and HB are extracted from a WB training data set. The HB spectral envelope is determined from the HB code vector whose corresponding NB code vector is closest in shape to the spectral envelope of the frame of input NB speech under analysis. Incorporating the voiced and unvoiced information in the CB mapping is performed by splitting the codebooks into voiced and unvoiced entries. This is performed by the calculation of a binary voicing indication to divide the training set into spectral envelopes from voiced and unvoiced frames. Each data set is then used to train the NB and HB codebooks resulting in two codebooks for the voiced frames and two codebooks for the unvoiced frames. Using the voicing information, achieved the smallest spectral distortion as compared to other conventional CB based methods implemented in [11].

Voiced sounds, sibilants and stop-consonants classes are used in [32] for HB spectral shape estimation. This algorithm is a non-model based BWE algorithm. The classification task is being accomplished by a deterministic finite-state machine in which the classification decision depends on the previous state and the feature vector extracted from the current speech frame. A rule-based classification approach uses a set of empirically determined decision rules and threshold values as criteria for the different classes. The thresholds are independent of other features and they have been set manually as a result of experiments on a training data-set. The classification procedure involves the following major steps. First, the classifier checks if the current frame meets the criteria of a sibilant sound. If a frame is not found to be a sibilant, the classifier tests if the frame satisfies the criteria of a plosive sound. If not, the frame falls into the category of voiced sounds. The shaping filter is constructed in the frequency domain using cubic-spline interpolation of five control points at 4, 5, 6, 7 and 8 kHz. The $l$ control point is calculated using a predefined formula of the form:

$$\mathbf{C}_l\left(k\right) = b_l + a_l \cdot \mathbf{x_{SS}}\left(k\right), \quad l = 1, ..., 5 \quad , \tag{3.5}$$

where $k$ is the frequency index, $b_l$ and $a_l$ are different control point constants calculated offline for each phonetically motivated category using a search based on the genetic algorithm [14]. $\mathbf{x_{SS}}$ is the NB signal analyzed spectral slope.

Phonetic transcription is used in [6] for supervised training of an HMM statistical model. The states of the HMM are defined by a vector quantization codebook (CB) of real HB spectral envelopes (represented in the cepstral domain) using the LBG algorithm. Each state $S_i\left(m\right),\ i=1,\ldots,N_s$, represents a speech phoneme spectral envelope, where $i$ is the state index, $N_s$ is the number of states and $m$ is the current frame time index. The HMM statistical model is trained offline using real WB data. The real state sequence is calculated first by calculating the HB spectral envelopes and classifying them to the pre-defined states CB. From the real state sequence, the state initial probability, $p\left(S_i\right)$, and the state transition probability of the Markov chain from state $j$ to state $i$, $p\left(S_i\left(m\right)\left|S_j\left(m-1\right)\right.\right)$, are calculated. Using the NB analyzed feature vector, $\mathbf{x}$, which represent each state, the observation probability for each state, $p\left(\mathbf{x}\left|S_i\right.\right)$, is calculated. This probability is approximated by GMM parameters with $N_g$ mixtures, which are estimated for each state by the expectation-maximization (EM) algorithm [27].

The state probabilities for an input speech frame are extracted from the a-posteriori probability density function (PDF). The observation sequence of the NB feature vector $\mathbf{x}$ up to the current frame is defined as $\mathbf{X}\left(m\right)=\left\{\mathbf{x}\left(1\right),\mathbf{x}\left(2\right),\ldots,\mathbf{x}\left(m\right)\right\}$. The conditional probability $p\left(S_i\left(m\right)\left|\mathbf{X}\left(m\right)\right.\right)$ expresses the a-posteriori probability. It is recursively calculated for each state by

$$\begin{array}{l} p\left(S_i\left(m\right)\left|\mathbf{X}\left(m\right)\right.\right)=C\cdot p\left(\mathbf{x}\left(m\right)\left|S_i\left(m\right)\right.\right)\cdot \\ \sum\limits_{j=1}^{N_s} p\left(S_i\left(m\right)\left|S_j\left(m-1\right)\right.\right)p\left(S_j\left(m-1\right)\left|\mathbf{X}\left(m-1\right)\right.\right)\quad , \end{array} \tag{3.6}$$

where $C$ is a normalization factor to allow all the state probabilities to sum up to one [6].

The CB of HB spectral envelope, $\hat{\mathbf{y}}$, and the calculated state a-posteriori probabilities are used to perform a minimum mean square error (MMSE) estimation of the HB spectral envelope:

$$\tilde{\mathbf{y}}=\sum_{i=1}^{N_s}\hat{\mathbf{y}}_{\mathbf{i}}\cdot p\left(S_i\left(m\right)\left|\mathbf{X}\left(m\right)\right.\right), \tag{3.7}$$

where $\tilde{\mathbf{y}}$ represents the estimated HB spectral envelope in the current analyzed speech frame.

The advantage in using the HMM model for HB spectral envelope estimation is in the fact that the HMM model describes a time evolving process. This yields information for the spectral envelope estimation both from the current frame extracted features (as in the CB mapping approach) and from past information.

### Estimation Based on Speaker Characteristic

Another way to improve the HB spectral envelope estimation is by making it robust to variation of speakers. Different speakers yield different formants locations even when representing the same speech linguistic content [35]. Using speaker related features such as vocal tract area function (VTAF) allows a better speaker-dependent estimation. The VTAF represents the vocal tract's physical shape as a function of the distance from the glottis. The concatenated tube model is used for VTAF shape representation [35, 42].

The algorithm in [13] uses this model to estimate the formants locations in the HB for voiced sounds. The estimated formants frequencies are used to construct a minimum-phase filter with a descending amplitude that is not specified in the paper. The constructed filter is used in the frequency domain to shape the generated excitation signal. The peaks at the two highest frequencies in the NB are used in the calculation of the placement of the synthetic formants. This is justified by the fact that these estimated resonance frequencies are the most likely to be resonances of the same front-most cavity. When this front-most cavity is considered to be a uniform tube, opened in the front and closed in the rear end, the resonance frequencies $f_{n_f}$ with index $n_f$ are:

$$f_{n_f} = \frac{2n_f - 1}{4} \cdot \frac{c}{l}, \quad n_f = 1, 2, 3, ... \tag{3.8}$$

where $c$ is the speed of sound and $l$ is the length of the uniform tube. The use of a uniform tube is a rough approximation. To calculate the frequencies of the HB resonances an estimation of the front most tube length and the resonances number, $n_f$, is necessary. Assuming the highest formants frequencies in the NB, $f_{n_1}$ and $f_{n_2}$ are two consecutive resonance frequencies of the front-most cavity, the resonance number associated with $f_{n_1}$ can be estimated by:

$$n_1' = \frac{f_{n_1} \cdot 2l}{c} + \frac{1}{2} \tag{3.9}$$

and

$$n_1'' = n_2 - 1 = \frac{f_{n_2} \cdot 2l}{c} - \frac{1}{2}. \tag{3.10}$$

Another estimate of the resonance number corresponding to $f_{n_1}$ is the average of the previous estimates:

$$\bar{n} = \text{round}\left(\frac{n_1 + n_2}{2}\right) = \text{round}\left((f_{n_1} + f_{n_2}) \cdot \frac{l}{c}\right). \tag{3.11}$$

Since the resonance number is an integer the estimate, $\bar{n}$, is approximated to the nearest integer. By inserting $\bar{n}$ in (3.8) the tube length can be derived for each segment as:

$$\frac{c}{l} = 2\left(f_{n_1} + f_{n_2}\right). \tag{3.12}$$

A shorter distance between the frequencies implies a longer tube. The fraction $\frac{c}{l}$ is limited, a maximum tube length of 20 cm is assumed which is a reasonable physically limit. The limitation results in a lowest distance limit between the resonance frequencies of 0.9 kHz. The $n_{f_{\mathrm{HB}}}$ synthetic HB formant frequencies are then calculated with (3.8) for $n = \bar{n} + 1 + k, \quad k = 1, ..., n_{f_{\mathrm{HB}}}$.

In [8, 25] the WB spectral envelope is obtained from an estimation of the WB VTAF shape. As speech is produced by a physical system modeled by the VTAF, the estimation of the WB VTAF shape is done in [8, 25] by interpolating the NB VTAF shape. Pre-emphasizing is first performed on the received NB speech signal to remove the effect of the lip radiation and the glottal wave shape from the vocal tract filter. The area coefficients representing the NB VTAF are calculated using the reflection coefficients extracted from the LPC. According to the constraint in equation (2.2), $N_A = 8$ area coefficients are calculated for the NB VTAF. The BWE algorithm goal is to extend the spectral envelope bandwidth by estimating $N_A = 16$ area coefficients that represent the WB VTAF shape. Based on the equal-area principle stated in [42], each uniform area section in the VTAF should have an area that is equal to the mean area of an underlying continuous area function of a physical vocal tract. Hence, doubling the number of sections corresponds to splitting each section into two in such a way that, preferably, the mean value of their areas equals the area of the original section. The estimation of the WB VTAF is performed by first interpolation the NB VTAF using cubic-spline interpolation. Then, resampling of the interpolated NB VTAF is performed in the points that correspond to the new sections centers. Because the resampling is performed at points that are shifted by 1/4 of the original points, this interpolation is called shifted-interpolation [8, 25]. An example of NB VTAF and an interpolated VTAF and the corresponding spectral envelopes is shown in Fig. 3.4.

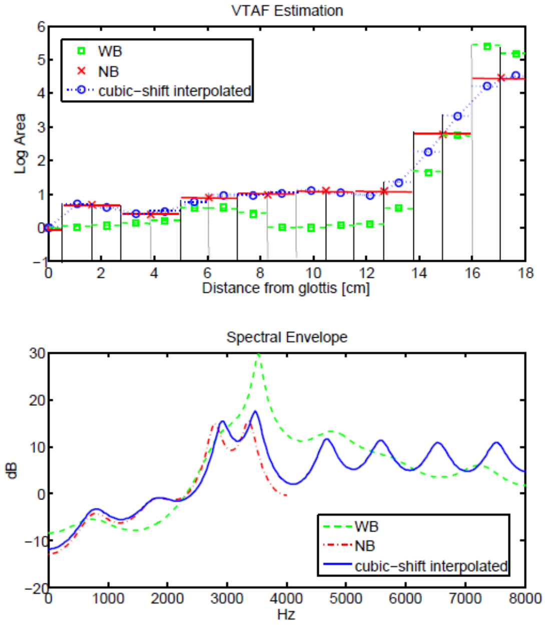


Figure 3.4: Example of original WB, NB and shifted interpolated NB VTAF and their corresponding spectral envelopes.

### 3.3.2 Highband Gain Estimation

Gain estimation is possible by WB spectral envelope estimation and gain adjustment to match the received NB spectral envelope gain, as in [22]. It is also possible to estimate the gain as part of the HB spectral envelope estimation, like in [36]. Gain estimation using gain adjustment to the NB speech signal gain is a common technique in BWE algorithms. In [22], the WB spectral envelope is estimated using CB mapping. The gain factor is calculated as the ratio between the power of the virtually re-synthesized NB signal, using the estimated WB spectral envelope, and the power of the real NB signal. This yields

$$g = \frac{\tilde{P}_{\mathrm{NB}}}{P_{\mathrm{NB}}} \quad , \tag{3.13}$$

where $g$ is the gain factor, $\tilde{P}_{\mathrm{NB}}$ is the power of the virtually re-synthesized NB signal and $P_{\mathrm{NB}}$ is the power of the real NB signal. Adjusting the BWE signal using the gain factor produces equal powers of the estimated BWE signal in the NB region and the received NB signal.

In [36], gain estimation is performed by linear mapping of the NB gain to the HB gain. The linear mapping coefficients are trained offline using a large speech database. A set of 60 HB spectral envelopes was built offline by partitioning a large amount of real HB spectral envelopes by their gain. Each entry of the set represents a different HB gain value with a difference of 1 dB between consecutive entries. The estimated HB gain is used to extract HB spectral envelope. Small changes in the estimated HB energy correspond to small changes in HB spectral envelope shapes. These small changes permit the explicit control of the time evolution of the HB spectral envelope shape by controlling the time evolution of the HB energy. Smooth evolution of the HB spectrum, at least within distinct speech segments, can be important for ensuring natural-sounding, high-quality output BWE speech with minimal artifacts.

## 3.4 Lowband Signal Extension

The missing LB frequencies reduce the speech naturalness. On a perceptual frequency scale, such as the Mel-scale, the LB covers approximately three Mel-bands and the HB covers approximately four Mel-bands. This yields that the LB is almost as wide as the HB on a perceptual scale. Few BWE algorithms deal with the missing LB frequencies as any minor estimation error in the LB can cause major artifacts to the BWE signal. The BWE algorithms that deal with LB extension can be divided into two types. Algorithms that use existing information in the LB frequencies after the processing with the IRS filter, described in Chapter 1, and algorithms that estimate the LB frequencies from information extracted from the NB frequencies.

LB equalization allows compensating for the IRS filter response and extracting the real missing LB magnitude information. In [33] an equalizer was designed to recover the lost signal in the LB. The equalizer has a boost of 10 dB at 100 Hz. The frequency response of the signal after the equalizer is almost flat from 100–300 Hz. The limitation of such algorithm is in the amplification of backgroung noise in the LB frequency range due to the equalization. This may add artifacts to the extended signal.

Estimation of the missing LB is performed in [13,21]. Fig. 3.5 describes the excitation extension in [21]. The NB excitation is upsampled to 16 kHz and new LB excitation component is obtained by the application of a quadratic function. The quadratic function generates an harmonic structure in the frequency domain with the pitch frequency as the fundamental frequency. The following stages include whitening of the generated LB

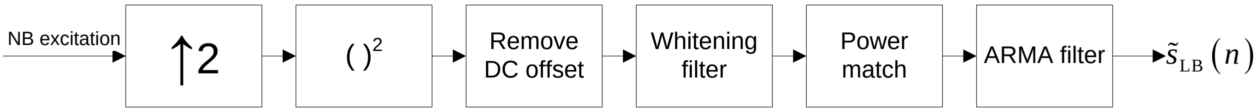


Figure 3.5: Block Diagram of LB signal extension (from [22]).

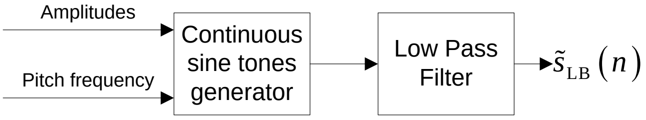


Figure 3.6: Block Diagram of LB extension (from [13]).

excitation and power matching to the NB signal. The calculated LB excitation is then filtered through a fixed auto-regressive moving-average (ARMA) filter with an adaptive amplification factor. The ARMA filter has two complex conjugate poles at 120 Hz and two complex conjugate poles at 215 Hz. These locations correspond to the mean male and female pitch frequencies, respectively.

The synthesis of the estimated LB signal in [13] is performed by generating and summing sine tones at the pitch frequency and its harmonics up to 300 Hz. The tones are generated with an amplitude equal to a fraction of the amplitude level of the first formant (see Fig. 3.6).

The limitation of algorithms that estimate the LB comes from the fact that the estimated LB signal can substantially mask higher frequencies when a high amplitude level is used. Masking denotes the phenomenon of one sound, the masker, making another sound, the masked one, inaudible. The risk of masking yields that extra caution must be taken when generating the estimated LB signal.

## 3.5 Summary

We have presented methods for BWE of NB speech. LB extension aims to improve speech naturalness and HB extension aims to improve speech quality as well as intelligibility. General BWE algorithms include spectral components generation, spectral components shaping, gain adjustment of shaped components to match the gain of the received NB signal, and synthesis of the estimated WB signal.

In order to improve BWE algorithms, the source-filter speech model is incorporated into the algorithm. This allows the separation of the BWE into two independent tasks. The extension of the excitation bandwidth and the extension of the spectral envelope. Further improvement of the BWE algorithm is achieved by incorporating speech sound class information in the estimation process. This can be performed by classification of the speech into few speech classes such as voiced and unvoiced, or by classifying the speech to the whole speech phonemes space. Another method to improve a BWE algorithm is by making it robust to variation of speakers. This can be performed by using speaker related features such as the VTAF and the usage of the concatenated tube model for speaker representation.

In the next chapter we will present a new BWE algorithm that is based on the speech source-filter model. This algorithm involves phonetic and speaker dependent estimation of the HB spectral envelope.

# Chapter 4

# Proposed Bandwidth Extension System

The proposed BWE algorithm is described in this chapter. Our BWE approach uses both phonetic and speaker dependent information for HB spectral envelope estimation. The first step employs an HMM model to classify each speech frame to a specific phoneme type. The second step finds a speaker specific WB spectral envelope by WB VTAF shape estimation from the calculated NB VTAF shape. The third step includes a new proposed postprocessing step, involving modification of the estimated WB VTAF, allows better gain adjustment and smoothing in time of the estimated spectral envelope.

The general BWE algorithm scheme is described in Fig. 4.1. The system can be divided into four stages. Stage I carries out preprocessing and feature extraction. The input to this stage is the received NB speech signal $s_{\mathrm{NB}}(n)$ with sample index $n$. The NB speech signal is framed, upsampled and equalized in the low frequencies. Three sets of feature vectors are extracted from each preprocessed frame of the received signal: Frequency-based features, $\mathbf{x_1}$, for speech-state estimation; NB VTAF feature vector, $\mathbf{x_2}$, for WB VTAF estimation, and NB excitation, $\mathbf{x_3}$, for WB excitation generation.

In Stage II of the algorithm, the estimation of the WB spectral envelope $\tilde{\phi}_{\mathrm{WB}}(k)$, with frequency index $k$, is performed. It is calculated in a three-step process. In the first step, speech state estimation yields the probability of being in a specific speech-phoneme related state. The WB VTAF, $\mathbf{\tilde{A}_{WB}}$, is then estimated, in the second step, from the calculated NB VTAF. Postprocessing of the estimated WB VTAF, in the third step, allows better

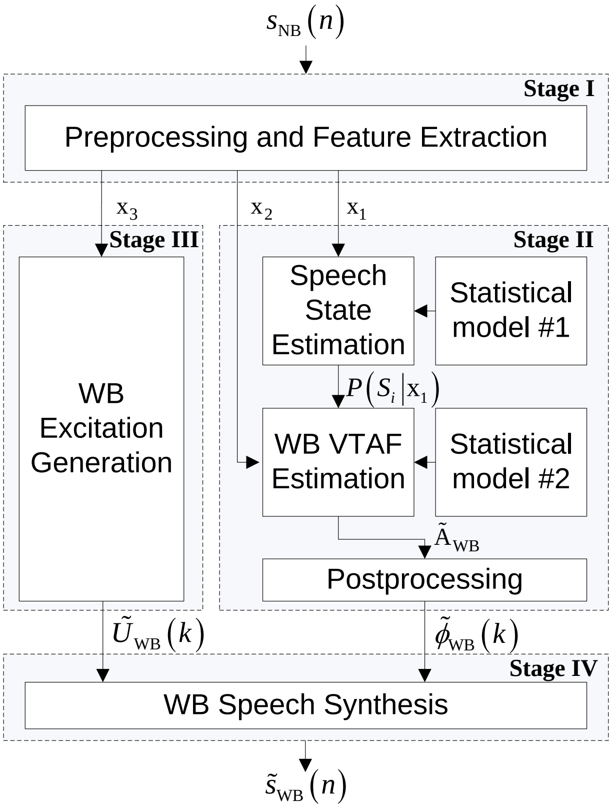


Figure 4.1: Block diagram of the proposed BWE algorithm.

gain adjustment and smoothing in time of the estimated WB spectral envelope.

In Stage III of the algorithm, the WB excitation, $\tilde{U}_{\text{WB}}(k)$ is generated. The HB excitation is generated using a simple spectral copy of the calculated NB excitation. In the last stage of the algorithm, Stage IV, the output WB speech signal $\tilde{s}_{\text{WB}}(n)$ is synthesized in the frequency domain, without changing the received NB signal.

This chapter is organized as follows. In Section 4.1, we describe the preprocessing and feature extraction stage. In Section 4.2, we present the WB spectral envelope estimation stage. In Section 4.3, we present the WB excitation generation. Finally, in Section 4.4, we describe the WB speech synthesis.

## 4.1 Preprocessing and Feature Extraction

The preprocessing and feature extraction block diagram is presented in Fig. 4.2. The received NB speech signal is segmented into frames of 10msec duration. Processing the

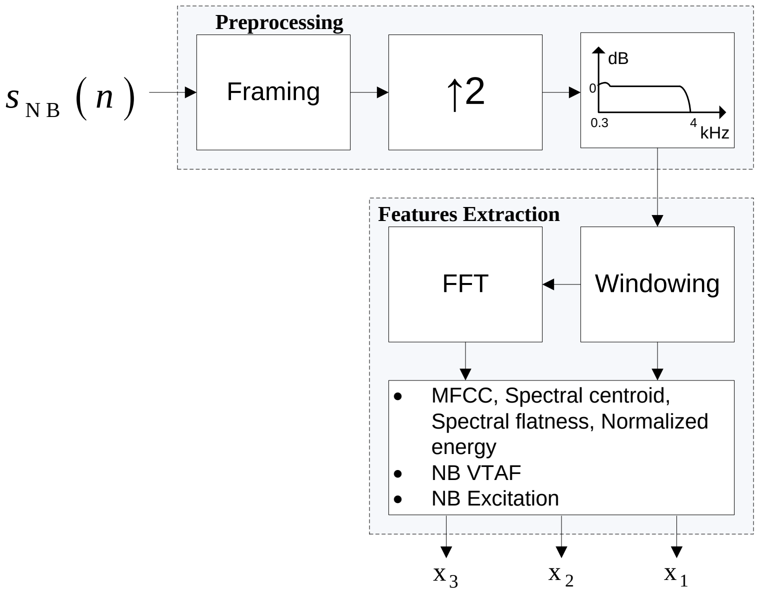


Figure 4.2: Block diagram of the preprocessing and feature extraction stage.

received NB speech signal in short frames will allow to implement a real-time application of the algorithm. The speech frame is upsampled to 16 kHz sampling rate and filtered through a LPF with 4 kHz cutoff frequency and 10dB boost at 300 Hz. The equalizer LPF frequency response is presented in Fig. 4.3. The 10dB boost equalizes a typical telephone channel filter response [4,15], which attenuates the speech signal at and below 300 Hz. This equalization adds naturalness to the NB signal. The equalized speech signal is segmented into frames of 20msec duration, with 10msec overlap between frames. Processing the speech signal in short frames with overlap allows to operate in the frequency domain [35]. The equalized frame is then windowed using a Hamming window. The Hamming window advantage is that it is simple to implement and has small spectral leakage. This allows better spectral analysis of the speech signal. The preprocessed frame is analyzed both in the time and the frequency domain. A fast Fourier transform (FFT) of length 512 is used for frequency domain frame calculation.

Three sets of features are extracted from the upsampled and equalized speech frame. The purpose of the first feature vector, $\mathbf{x_1}$, is to allow good separation of different speech classes that give different HB spectral envelope shapes [19]. All of the $\mathbf{x_1}$ features are frequency domain based features. The analysis in the frequency domain allows control of the relevant signal bands for the features extraction. The feature vector $\mathbf{x_1} \in \mathbb{R}^{13}$ consists

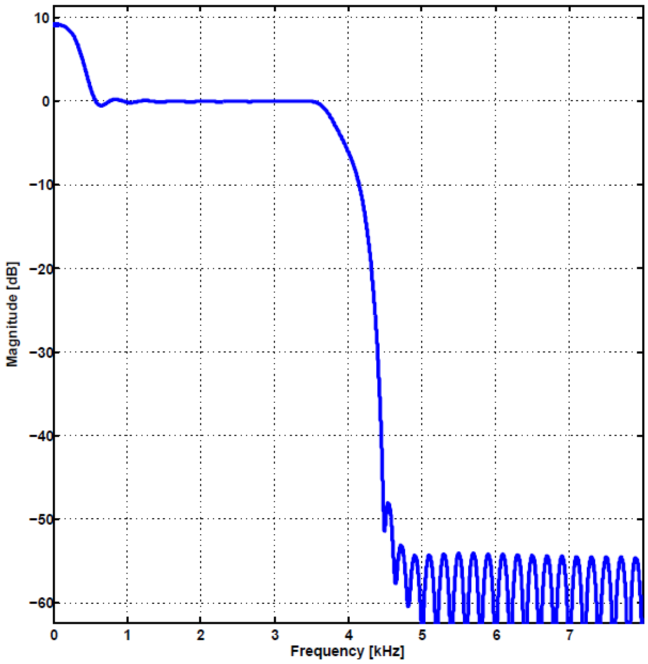


Figure 4.3: Frequency response of the equalizer lowpass filter (see filter coefficients at Appendix B).

of the following features (which were discussed in Section 2.2):

- *Mel Frequency Cepstral Coefficients* (MFCC) of nine subbands from 300 to 3400Hz. The MFCC are commonly used in speech recognition algorithms. They were shown to have high NB to HB speech mutual information and to provide good class separation [30].

- *Spectral centroid* of the NB power spectrum, which is generally high for unvoiced sounds [31].

- A *spectral flatness* measure [31] that indicates the tonality of the speech signal.

- *Spectral slope*, which is useful for discriminating voiced frames from sibilants and plosives [32].

- *Normalized frame energy* [20].

The second feature vector, $\mathbf{x_2}$, contains the area coefficients that represent the speaker's NB VTAF shape. The area coefficients are calculated from the reflection coefficients as described in (2.3). Since the preprocessed NB frame is upsampled to 16 kHz, we use $N_A = 16$ area coefficients for NB VTAF calculation.

The last extracted feature vector, $\mathbf{x_3}$, is the NB excitation, which is calculated in the frequency domain by dividing the NB signal spectrum by the NB spectral envelope signal. A fast Fourier transform (FFT) of length 512 is used for frequency domain signals calculation from the upsampled frame.

## 4.2 Wideband Spectral Envelope Estimation

The estimation of the WB spectral envelope is carried out in three steps. In the first step, the speech state which represents a specific speech phoneme is estimated using an HMM-based statistical model. The second step consists of estimating the specific speaker WB VTAF shape by a codebook search, using the speaker calculated NB VTAF shape. Postprocessing of the estimated WB VTAF is conducted in the last step, to reduce possible artifacts due to estimation errors in the previous steps.

### 4.2.1 Speech State Estimation

The HMM statistical model was trained offline using the TIMIT transcription. Each frame was associated with a state $S_i\left(m\right),\ i = 1, \ldots, N_s$, which represents a speech phoneme (one state for each phoneme), where $i$ is the state index, $N_s$ is the number of states and $m$ is the current frame time index. From the training database, the state $S_i$ and the first feature vector $\mathbf{x_1}$ of each speech frame were extracted. The following probability density functions (PDFs) were calculated:

- $p\left(S_i\right)$ - Initial probability of each state.
- $p\left(S_i\left(m\right)\left|S_j\left(m-1\right)\right.\right)$ - Transition probability of the first order Markov chain from state $j$ to state $i$.
- $p\left(\mathbf{x_1}\left|S_i\right.\right)$ - Observation probability for each state. This probability is approximated by GMM parameters with $N_g$ mixtures, which are estimated for each state by the expectation-maximization (EM) algorithm [20].

The state initial probability and transition probability are necessary to describe the process of speech production which can be assumed to be short-term stationary. The state sequence is assumed to be governed by a first-order Markov chain. The dependencies between the HMM states of consecutive signal frames is considered in the HMM statistical model.

The observation probability is defined as the conditional PDF of the first feature vector $\mathbf{x_1}$. The conditioning is with respect to the HMM state $S_i$ such that there is a separate PDF $p\left(\mathbf{x_1}\left|S_i\right.\right)$ of the dimension of $\mathbf{x_1}$ for each state. According to the definition of the HMM, it is assumed that the observation $\mathbf{x_1}\left(m\right)$ for each frame only depends on the state of the Markov chain during that particular frame. The high dimension PDF, $p\left(\mathbf{x_1}\left|S_i\right.\right)$, is approximated by GMM parameters as:

$$p\left(\mathbf{x_1}\left|S_i\right.\right) = \sum_{l=1}^{N_g} \rho_{il} G\left(\mathbf{x_1}; \boldsymbol{\mu}_{il}, \mathbf{V}_{il}\right) \ , \tag{4.1}$$

where $G\left(\mathbf{x_1}; \boldsymbol{\mu}_{il}, \mathbf{V_{il}}\right)$ is the $l$th Gaussian mixture component with scalar weighting factor $\rho_{il}$, mean vector $\boldsymbol{\mu}_{il}$ and covariance matrix $\mathbf{V}_{il}$. For the training of the Gaussian mixture model the iterative EM algorithm is used [20].

In the application phase, the state probabilities for an input speech frame are extracted from the a-posteriori PDF. We denote the observation sequence of the first feature vector $\mathbf{x_1}$ up to the current frame as $\mathbf{X_1}\left(m\right) = \left\{\mathbf{x_1}\left(1\right), \mathbf{x_1}\left(2\right), \ldots, \mathbf{x_1}\left(m\right)\right\}$. The conditional probability $p\left(S_i\left(m\right)\left|\mathbf{X}_1\left(m\right)\right.\right)$ expresses the a-posteriori probability. It is recursively calculated for each state by

$$\begin{gathered} p\left(S_i\left(m\right)\left|\mathbf{X}_1\left(m\right)\right.\right) = C_1 \cdot p\left(\mathbf{x}_1\left(m\right)\left|S_i\left(m\right)\right.\right) \cdot \\ \sum_{j=1}^{N_s} p\left(S_i\left(m\right)\left|S_j\left(m-1\right)\right.\right) p\left(S_j\left(m-1\right)\left|\mathbf{X}_1\left(m-1\right)\right.\right) \ \ , \end{gathered} \tag{4.2}$$

where $C_1$ is a normalization factor to allow all the state probabilities to sum up to one [6]. Choosing the state with the highest a-posteriori probability yields a hard-decision for the current speech frame linguistic content.

### 4.2.2 Wideband VTAF Estimation

Now, we wish to estimate a suitable WB spectral envelope for the estimated speech state. For this purpose we estimate the speaker's VTAF shape. As the VTAF shape models the physical speech production system, we wish to find the closest WB VTAF shape to the calculated speaker's NB VTAF shape. We use a second statistical model that incorporates a set of WB VTAF codebooks (CBs). For each of the $N_s$ states, we have a CB with $N_{\mathrm{CB}}$ entries. The CBs were trained offline with real WB VTAF data, extracted from the TIMIT train database, using the Linde, Buzo, Gray training (LBG) algorithm [15]. We denote the calculated NB VTAF as $\mathbf{A_{NB}}$ and the CB entries corresponding to the estimated state $S_i$ as $\mathbf{A}^{S_i}_{\mathbf{WB}}(j)\,,\ j = 1, \ldots, N_{\mathrm{CB}}$. The optimal WB VTAF $\tilde{\mathbf{A}}^{S_i}_{\mathbf{WB}}$ for the estimated state in frame $m$ is picked by minimizing the Euclidean distance between $\mathbf{A_{NB}}$ and $\mathbf{A}^{S_i}_{\mathbf{WB}}(j)\,,\ j = 1, \ldots, N_{\mathrm{CB}}$ :

$$\begin{aligned} \tilde{\mathbf{A}}^{S_i}_{\mathbf{WB}} &= \mathbf{A}^{S_i}_{\mathbf{WB}}\left(j^{opt}\right), \\ j^{opt} &= \arg\min_{j=1}^{N_{\mathrm{CB}}} \left\| log\left(\mathbf{A_{NB}}(m)\right) - log\left(\mathbf{A}^{S_i}_{\mathbf{WB}}(j)\right)\right\|_2^2 . \end{aligned} \tag{4.3}$$

In-order to reduce artifacts due to erroneous state estimation, we use $N_{\mathrm{best}}$ states with the highest a-posteriori probabilities $p_1, \ldots, p_{N_{\mathrm{best}}}$ for WB VTAF estimation

$$\tilde{\mathbf{A}}_{\mathbf{WB}} = C_2 \cdot \left(p_1 \cdot \tilde{\mathbf{A}}^{S_{i_1}}_{\mathbf{WB}} + \ldots + p_{N_{\mathrm{best}}} \cdot \tilde{\mathbf{A}}^{S_{i_{N_{\mathrm{best}}}}}_{\mathbf{WB}}\right), \tag{4.4}$$

where $C_2$ is a normalization factor to constrain the highest $N_{\mathrm{best}}$ probabilities to sum up to one.

### 4.2.3 Postprocessing

After the last step that was described in Subsection 4.2.2, we have an initial WB VTAF estimation from Eq. (4.4), $\tilde{\mathbf{A}}^{\mathbf{0}}_{\mathbf{WB}} = \tilde{\mathbf{A}}_{\mathbf{WB}}$, where the superscript 0 indicate an initial estimation. This estimated WB VTAF is further processed to allow better spectral envelope gain adjustment and smoothing in time. Better gain adjustment can be achieved by better fitting the lower band of the estimated WB spectral envelope to the calculated NB spectral envelope. Better smoothness in time can be achieved by reducing time discontinuities of estimated WB spectral envelopes. The postprocessing step is fully described in the remainder of this sub-section. Fig. 4.4 presents the block diagram of the proposed postprocessing.

We denote the formant frequencies of the NB and the estimated WB spectral envelopes by $\mathbf{f_{NB}}$ and $\tilde{\mathbf{f}}_{\mathbf{WB}}$, respectively. The shape fitting of the estimated WB spectral envelope is conducted by tuning the lower subset of $\tilde{\mathbf{f}}_{\mathbf{WB}}$ to $\mathbf{f}_{\mathbf{NB}}$. The tuning is done iteratively by perturbing the WB VTAF area coefficients [40]. The iterative tuning process is conducted only in voiced speech frames, as those frames are characterized by strong NB formant frequencies.

The VTAF is perturbed by using a sensitivity function. The sensitivity function relates small changes in VTAF to changes in formant frequencies. We denote the VTAF values by $A_{n_A},\ n_A = 1, \ldots, N_A$, where $N_A$ is the number of area coefficients. The spectral envelope formant frequencies are denoted by $f_{n_f},\ n_f = 1, \ldots, N_f$, where $N_f$ is the number of formant frequencies. The sensitivity function $S_{n_f,n_A}$ relates a small change in $f_{n_f}$ to incremental changes in the area coefficients, via:

$$\frac{\Delta f_{n_f}}{f_{n_f}} = \sum_{n_A=1}^{N_A} S_{n_f,n_A} \frac{\Delta A_{n_A}}{A_{n_A}} \quad . \tag{4.5}$$

Here we set $\Delta f_{n_f}$ to be the difference between the desired formant frequency and the current formant frequency. Thus, $\Delta A_{n_A}$ is the needed perturbation in the value of area coefficient number $n_A$. A vector form of (4.5) is:

$$\begin{gathered}\mathbf{\Delta}\hat{\mathbf{f}}_{\left[N_f x 1\right]} = \mathbf{S}_{\left[N_f x N_A\right]} \cdot \mathbf{\Delta}\hat{\mathbf{A}}_{[N_A x 1]}, \\ \mathbf{\Delta}\hat{\mathbf{f}} \triangleq \left[\frac{\Delta f_1}{f_1}, \ldots, \frac{\Delta f_{N_f}}{f_{N_f}}\right]^T, \ \mathbf{\Delta}\hat{\mathbf{A}} \triangleq \left[\frac{\Delta A_1}{A_1}, \ldots, \frac{\Delta A_{N_A}}{A_{N_A}}\right]^T. \end{gathered} \tag{4.6}$$

The sensitivity function is calculated by measuring the formants frequencies deviation due to small area changes using (4.5).

The goal of each iteration is to minimize the difference between the calculated and estimated NB formant frequencies. The formant frequencies are obtained by spectral envelope peak picking. The VTAF perturbation is solved from (4.6) by:

$$\mathbf{\Delta}\hat{\mathbf{A}}_{[N_A x 1]} = \mathbf{S}^{\dagger}_{\left[N_A x N_f\right]} \cdot \mathbf{\Delta}\hat{\mathbf{f}}_{\left[N_f x 1\right]} \quad , \tag{4.7}$$

where $\mathbf{S}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{S}$. This solution minimizes the $\ell^2$ norm of $\mathbf{\Delta}\hat{\mathbf{A}}$ when $N_f < N_A$. This criterion allows minimal area changes that approximate the desired formant frequencies changes. The pseudo-inverse of the sensitivity function matrix is calculated using the singular value decomposition (SVD) technique. Once $\mathbf{\Delta}\hat{\mathbf{A}}$ is calculated, the perturbation size for each VTAF area coefficient is $\Delta A_{n_A} = \Delta \hat{A}_{n_A} \cdot A_{n_A}$.

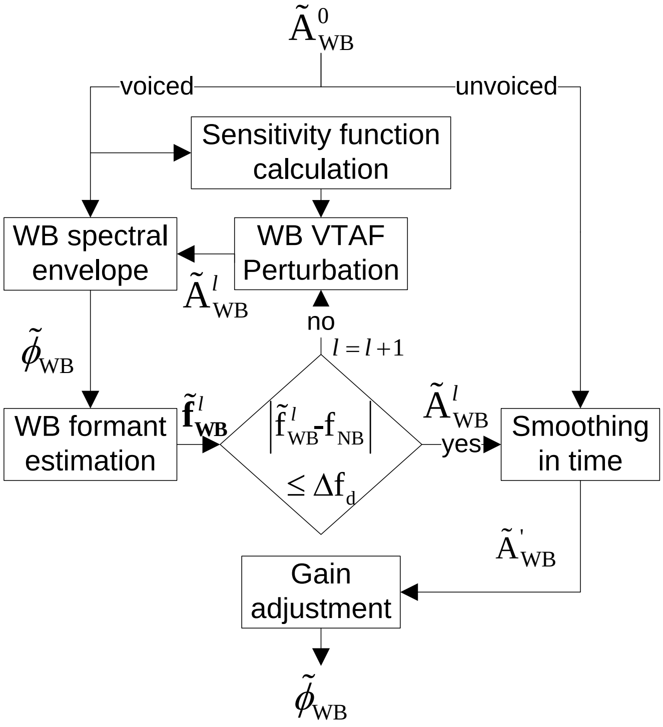


Figure 4.4: Block diagram of the proposed postprocessing step.

A new estimate of the WB VTAF is obtained by:

$$\tilde{\mathbf{A}}^{l+1}_{\mathbf{WB}}= \tilde{\mathbf{A}}^{l}_{\mathbf{WB}}+\boldsymbol{\Delta}\tilde{\mathbf{A}}^{l}_{\mathbf{WB}} \quad , \tag{4.8}$$

where $l$ is the iteration number and $\boldsymbol{\Delta}\tilde{\mathbf{A}}_{\mathbf{WB}} = [\Delta A_1, \ldots, \Delta A_{N_A}]^T$.

The stopping condition for the iterative process is the reaching of an allowed deviation, $\boldsymbol{\Delta}\mathbf{f_d}$, between $\mathbf{f_{NB}}$ and the corresponding lower subset of $\tilde{\mathbf{f}}_{\mathbf{WB}}$. No improvement in the frequencies deviation may imply a convergence problem and a large estimation error of the spectral shape. Hence, the estimated WB VTAF is updated only when the average frequencies deviations in the current iteration is smaller than that of the previous update. On average, 3.6 iterations were performed for each processed frame using $\boldsymbol{\Delta}\mathbf{f_d} = 50$ Hz. About 30% of the frames were processed using only one iteration and more than 75% of the frames were processed using four or less iterations.

When the described iterative process is finished, the result is a WB spectral envelope that its lower band is closer to the NB spectral envelope in terms of its NB formant locations compared to the initial WB spectral envelope calculated from $\tilde{\mathbf{A}}^{\mathbf{0}}_{\mathbf{WB}}$. Now, the estimated WB VTAF shape should be further processed to reduce possible artifacts and further improve the speech quality. This is done by smoothing in time and gain adjustment of the final estimated WB VTAF.

Smoothing in time is performed on the estimated WB VTAF under the assumption of physical continuity of vocal tract shape in time. Smoothing is done recursively by:

$$\tilde{\mathbf{A}}^{\prime}_{\mathbf{WB}}\left(m\right)=\beta\cdot\tilde{\mathbf{A}}^{\prime}_{\mathbf{WB}}\left(m-1\right)+\left(1-\beta\right)\cdot\tilde{\mathbf{A}}_{\mathbf{WB}}\left(m\right), \tag{4.9}$$

where $\beta = 0.7$ for voiced frames and $\beta = 0.5$ for unvoiced frames.

Gain adjustment is performed by first converting the smoothed estimate of the WB VTAF to a WB spectral envelope, as described in [35]. The calculated WB spectral envelope can now be gain adjusted to match the energy of the input NB spectral envelope in its lower band [22]. The HB spectral envelope $\tilde{\phi}_{hb}\left(k,m\right)$, is estimated for each processed frame in the frequency domain, where $m$ is the frame time index, and $k$ is the frequency bin index.

## 4.3 Wideband Excitation Generation

HB excitation generation is based on spectral copying of the NB excitation. The NB excitation in the transition band between 1.4-3.4 kHz is used repeatedly to fill the HB missing frequencies. This simple method allows keeping the original NB excitation signal untouched and filling all the missing HB frequencies with an excitation signal without any gap.

The drawback of this method is its lack of keeping the harmonic structure of the HB excitation for voiced frames. Informal listening tests showed that, in case of using real WB spectral envelope, the lack of keeping the generated excitation harmonic structure above 3.4 kHz does not significantly degrade the subjective quality of the extended speech signal. In addition, the pitch-synchronize modulation approach yield high computational complexity (as described in Chapter 3). Hence, a good compromise between subjective quality and computational complexity is using the spectral copy of NB excitation to the HB frequency range.

The HB excitation $\tilde{U}_{hb}\left(k,m\right)$, is generated for each processed frame in the frequency domain, where $m$ is the frame time index, and $k$ is the frequency bin index.

## 4.4 Wideband Speech Synthesis

The estimated final HB spectral envelope is used to shape the generated HB excitation in the frequency domain to receive the estimated HB speech signal, $\tilde{S}_{hb}\left(k,m\right)$, in the frequency domain.

$$\tilde{S}_{hb}\left(k,m\right)=\tilde{U}_{hb}\left(k,m\right)\cdot\tilde{\phi}_{hb}\left(k,m\right)\quad , \tag{4.10}$$

where $k$ is the frequency index and $m$ is the frame time index. This provides a HB speech component that is then concatenated in the frequency domain to the original NB signal to create the estimated WB signal.

$$\tilde{S}_{wb}\left(k,m\right)=\begin{cases} S_{nb}\left(k,m\right),0<k<3.4kHz \\ \tilde{S}_{hb}\left(k,m\right),3.4<k<8kHz \end{cases} \quad . \tag{4.11}$$

The time-domain speech frame is calculated from the obtained BWE signal transform using the inverse fast Fourier transform (IFFT). Two sequential time frames are combined by the overlap-add method using a Hann synthesis window. The advantage of the Hann window in speech synthesis is that it terminates with zero values on both ends, which allows smoother transition in time between the calculated speech frames. A synthesis of a speech frame is demonstrated in Fig. 4.5

## 4.5 Summary

We have presented the proposed BWE algorithm. The algorithm is divided into four stages. Stage I has two objectives. The first objective is to preprocess the received speech frame. This allows more accurate calculation of the speech features and improves the speech naturalness by equalizing the LB. The second objective of Stage I is to extract the speech features that would be used in the next algorithm stages. The extracted features from Stage I are used in Stage II to estimate the HB spectral envelope. Stage II includes three steps: the estimation of the speech phonetic content using an HMM statistical model; the estimation of the WB spectral envelope using a WB VTAF codebook search for a best match to the calculated NB VTAF, and the postprocessing step which aims to reduce artifacts due to erroneous estimation in the first two steps and reduce possible

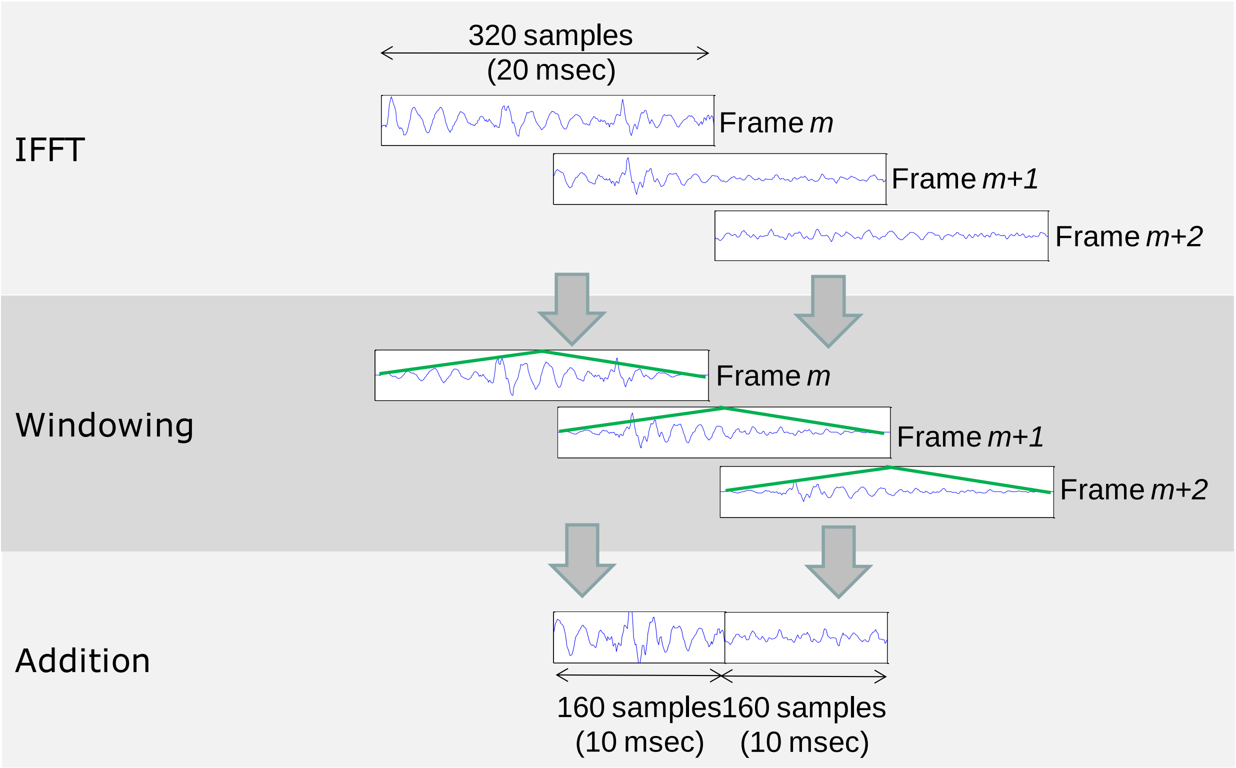


Figure 4.5: WB speech synthesis.

artifacts by better gain adjustment to the received NB speech. The generation of the WB excitation is conducted in Stage III using a simple spectral copy technique. The generated excitation with the estimated spectral envelope is used in Stage IV to synthesize the estimated WB speech signal using the overlap-add synthesis technique.

In the next chapter we present the experimental evaluation results of the proposed algorithm. The algorithm evaluation is conducted using both objective and subjective experiments.

# Chapter 5

# Experimental Results

Speech quality is a multi-dimensional term and its evaluation includes several factors. One factor that speech quality is composed of is the speech intelligibility. It is dependent on the speaker, the transmission channel (which includes the signal processing procedures the speech is subject to including the BWE), and the listener. Speech naturalness is another factor that depends on the transmission channel. Listening and conversational efforts are affected by the transmission channel, but are perceived subjectively different for each user.

Objective and subjective measures are commonly used to evaluate speech quality of speech processing algorithms. Subjective listening tests, which are carried out with people, take into account both human and technical influences. Testing scenarios would give different results for different listeners that would pay attention to different auditory events. Objective assessment methods, which are based on the human auditory model, take into account only the system influence on the speech quality. These measurements will give similar results for the same tested scenario. However, usually, these measurements are not highly correlated with the subjective tests results and hence, can't give accurate indication of the perceived speech quality.

In this chapter we evaluate the proposed BWE algorithm performance. We examine both objective and subjective measurements to evaluate the algorithm strength and needed future work.

## 5.1 Speech Quality Measurement

### 5.1.1 Objective Measurement

The primary selected objective measurement in the literature is the log spectral distance (LSD) of the HB power spectrums. It allows easy comparison to other BWE algorithms reported in the literature. The LSD is calculated for the $m^{th}$ frame by:

$$\mathrm{LSD}_m = \sqrt{\frac{1}{k_{\mathrm{high}} - k_{\mathrm{low}} + 1} \sum_{k=k_{\mathrm{low}}}^{k_{\mathrm{high}}} \left[ 10 \log_{10} \frac{P_m(k)}{\tilde{P}_m(k)} \right]^2} \quad , \tag{5.1}$$

where $P_m$ is the power spectrum of the original WB frame, and $\tilde{P}_m$ is the power spectrum of the corresponding BWE frame. The distortion is calculated using the FFT bin indices from $k_{\mathrm{low}}$ to $k_{\mathrm{high}}$.

The spectral distortion measure (SDM) [15] is also used in this research to evaluate the postprocessing step ability to reduce possible artifacts. This measure is a nonsymmetric weighted LSD. The distortion is generally calculated by using a decaying exponential for increasing frequencies and by giving higher penalty for spectral over-estimation than for under-estimation. The SDM measure for the $m^{th}$ frame is calculated by:

$$\mathrm{SDM}_m = \frac{1}{k_{\mathrm{high}} - k_{\mathrm{low}} + 1} \sum_{k=k_{\mathrm{low}}}^{k_{\mathrm{high}}} \xi_m(k) \quad , \tag{5.2}$$

where the distortion is calculated using the FFT bin indices from $k_{\mathrm{low}}$ to $k_{\mathrm{high}}$ and $\xi_m(k)$ is calculated as:

$$\xi_m(k) = \begin{cases} \Delta_m(k) \cdot \exp\{\alpha \Delta_m(k) - \beta k\} & , \text{if } \Delta_m(k) \geq 0 \\ \ln(-\Delta_m(k) + 1) \cdot \exp\{-\beta k\} & , \text{else} \end{cases}$$
$$\Delta_m(k) = 10 \log_{10} \tfrac{\tilde{P}_m(k)}{P_m(k)}.$$

Where $\alpha$ and $\beta$ are the weighting factors, $P_m$ is the power spectrum of the original WB frame, and $\tilde{P}_m$ is the power spectrum of the corresponding BWE frame. Fig. 5.1 shows the characteristic of the above defined distortion measure for $\alpha = 0.1$ and $\beta = 5$.

The motivation for using this measurement is the fact that high frequencies distortion is less significant for human perception. Another important issue that this measure deals with is giving more weight in computing the distortion to estimated spectrum that is above

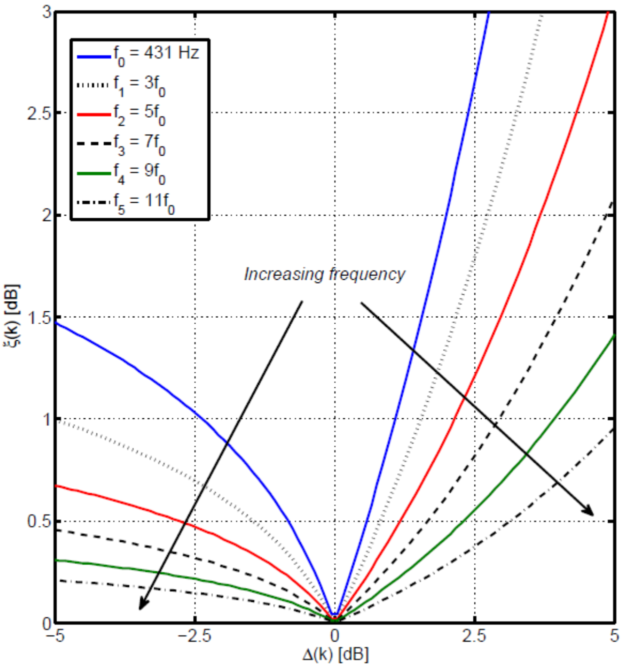


Figure 5.1: Frequency dependent branches of SDM for $\alpha = 0.1$ and $\beta = 5$ (from [15]).

the magnitude of the original one. Overestimated HB energy leads to undesirable audible artifacts that in the opposite case (of underestimation) does not cause any artifacts [28].

The last used objective measure is the formant frequencies error between HB estimated formant frequencies and their original counterparts. The formant frequencies are derived by peak picking in the spectral envelope. The error is calculated for each estimated HB formant, $\tilde{f}_{\mathrm{HB}}$, as:

$$Error = \left| \tilde{f}_{\mathrm{HB}} - f_{\mathrm{HB}} \right| , \tag{5.3}$$

where $\tilde{f}_{\mathrm{HB}}$ is the HB formant of the BWE estimated spectral envelope and $f_{\mathrm{HB}}$ is the corresponding HB formant of the original WB signal.

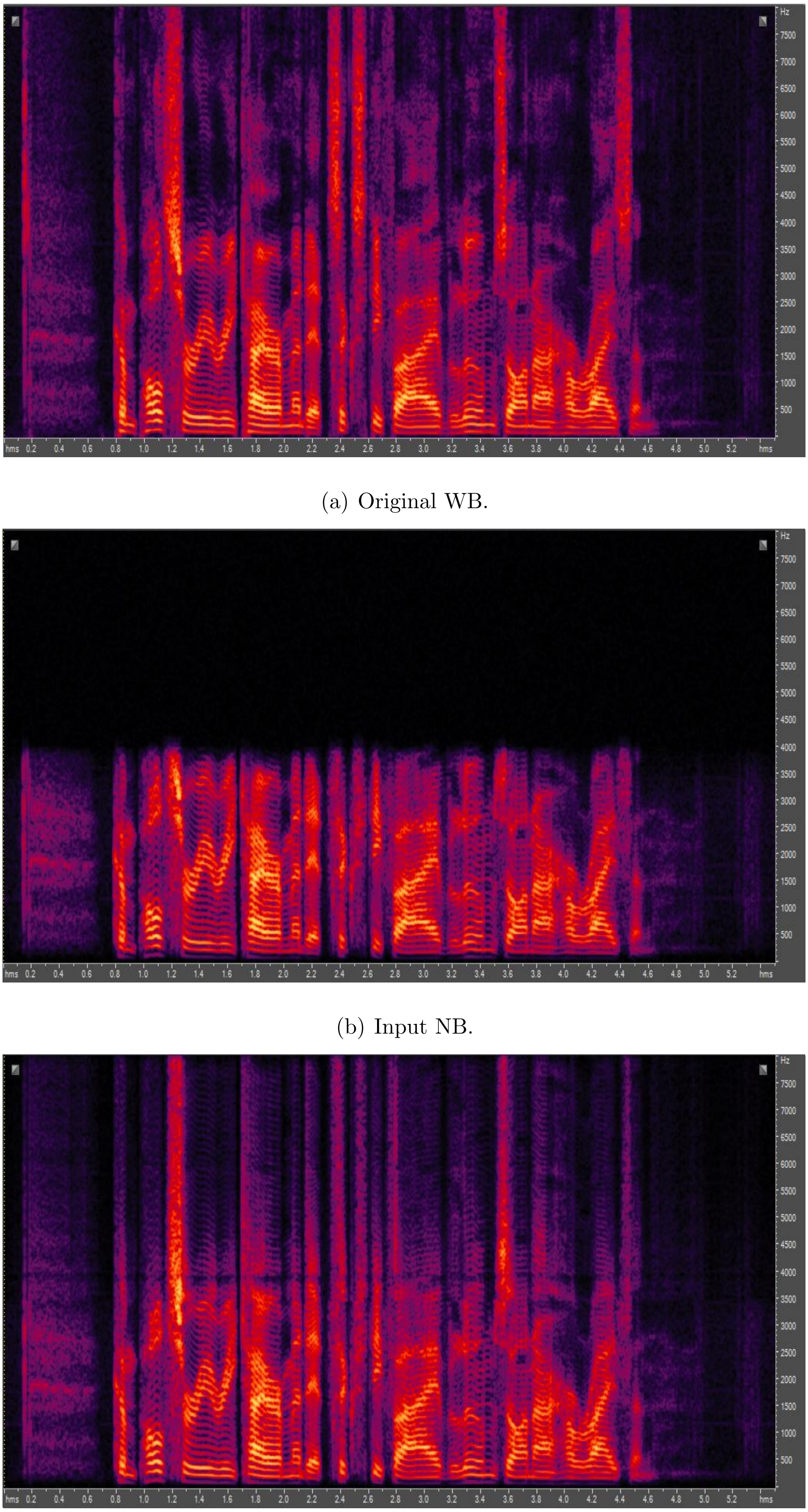
### 5.1.2 Subjective Measurement

Subjective listening tests are recognized as being the most reliable way of measuring speech quality. The mean opinion score (MOS) is probably the most common used subjective test to measure speech quality. In this test, each subject grades a test sentence by its quality from 1 which stands for bad quality to 5 which stands for excellent quality. The chosen subjective measure in this research is the multistimulus test with hidden reference and anchors (MUSHRA) test [1]. In this test the subject grades the processed speech test sentences in comparison to a reference sentence. The test sentences include all the sentences under test, an anchor sentence which should have the lowest quality compared to the reference sentence, and a hidden reference sentence similar to the reference sentence. This hidden reference is used for post-screening of subjects that gave low grade to the hidden reference. All the test sentences can be replayed by the subject at will. The main advantage of the MUSHRA test over the MOS test is that it is easier to perform, as it requires fewer participants to obtain statistically significant results [41]. This test also allows finer measurements of small differences because of the 0-100 score scale.

## 5.2 Performance Evaluation

To evaluate the algorithm performance, objective and subjective quality measurements were used. The proposed algorithm was implemented using Matlab. The following parameters were used: number of states $N_s = 61$ (symbols in the TIMIT lexicon), number of Gaussians per state $N_g = 16$ (as in [20]), number of CB entries per state $N_{\text{CB}} = 16$, number of VTAF area coefficients $N_A = 16$ (as in [25]), and number of states for VTAF estimation $N_{\text{best}} = 5$. The TIMIT WB training database, including 4620 sentences, was used for training both the HMM and the CB statistical models. The TIMIT WB test database, including 1680 sentences, was used as an input to the proposed algorithm after being preprocessed by a telephone channel filter and down sampled to 8 kHz. From the BWE processed signals and their original WB counterparts the following presented results and quality measurements were computed.

Fig. 5.2 and Fig. 5.3 depict the spectrograms of a male and a female utterances in the test database, showing the evolution of the speech spectra for the speech under different conditions. It can be seen from the spectrograms that the algorithm can estimate the HB spectrum with reasonable success. To further evaluate the system, results from the objective and subjective tests are analyzed.

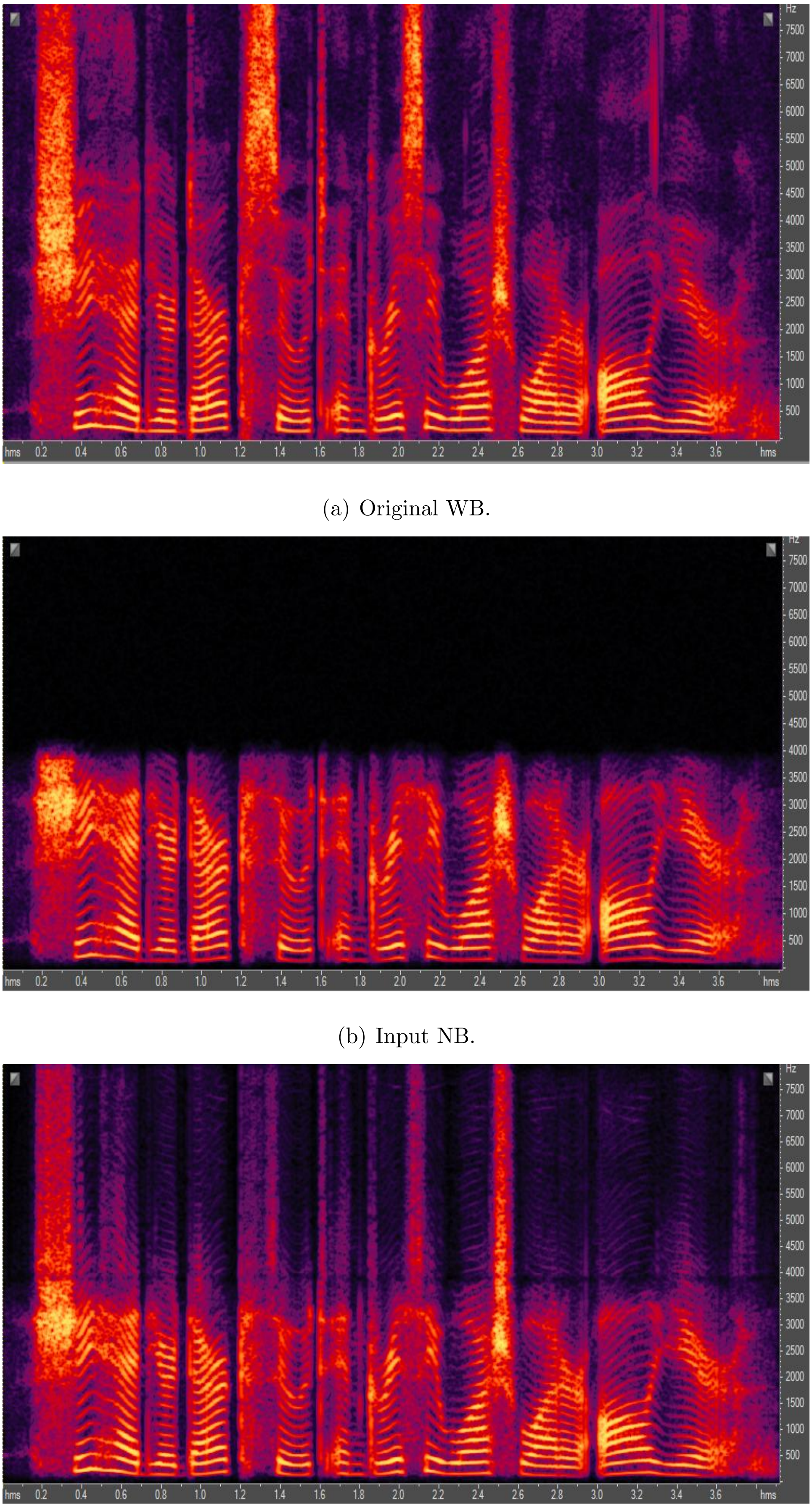


(a) Original WB.

(b) Input NB.

(c) Proposed BWE.

Figure 5.2: Spectograms of male utterances.



(a) Original WB.

(b) Input NB.

(c) Proposed BWE.

Figure 5.3: Spectograms of female utterances.

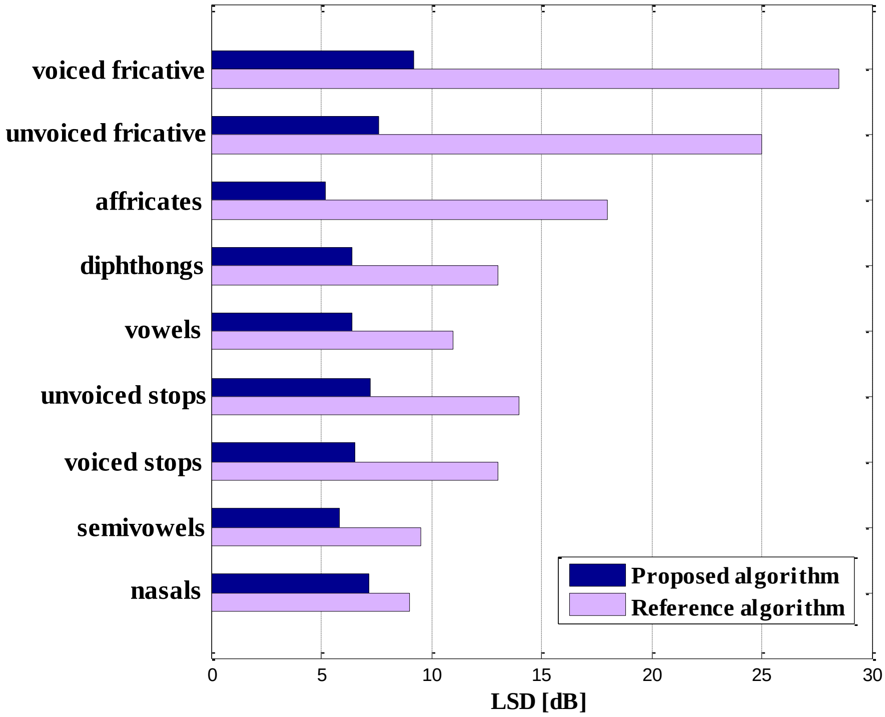


Figure 5.4: Average log spectral distortion for different phoneme categories.

### 5.2.1 Objective Evaluation

The first examined criterion was the LSD measure in different phonetic categories. The LSD is calculated as in Eq. (5.1). The distortion was calculated using the FFT bin indices from $k_{\mathrm{low}}$ to $k_{\mathrm{high}}$, corresponding to the frequency range from 4 to 7 kHz. For comparison reason with [32], the analysis was performed in frames of 256 samples (16 ms) using Hamming windowing, with 50% overlap between successive frames, and an FFT of length 1024.

Our results, in terms of the average LSD over phonemes in a given class, are compared in Fig. 5.4 to the results obtained in [32]. The X axis represents the LSD in dB, the Y axis represent the different phonemes categories. The results show improved performance of the proposed algorithm for all phoneme classes. A major improvement is obtained for fricative sounds. The LSD of Vowels is reduced by the proposed algorithm by 5 dB. The results demonstrate the effectiveness of phoneme dependent estimation of BWE speech frames.

The second evaluation criterion was the formant frequencies error between HB esti-

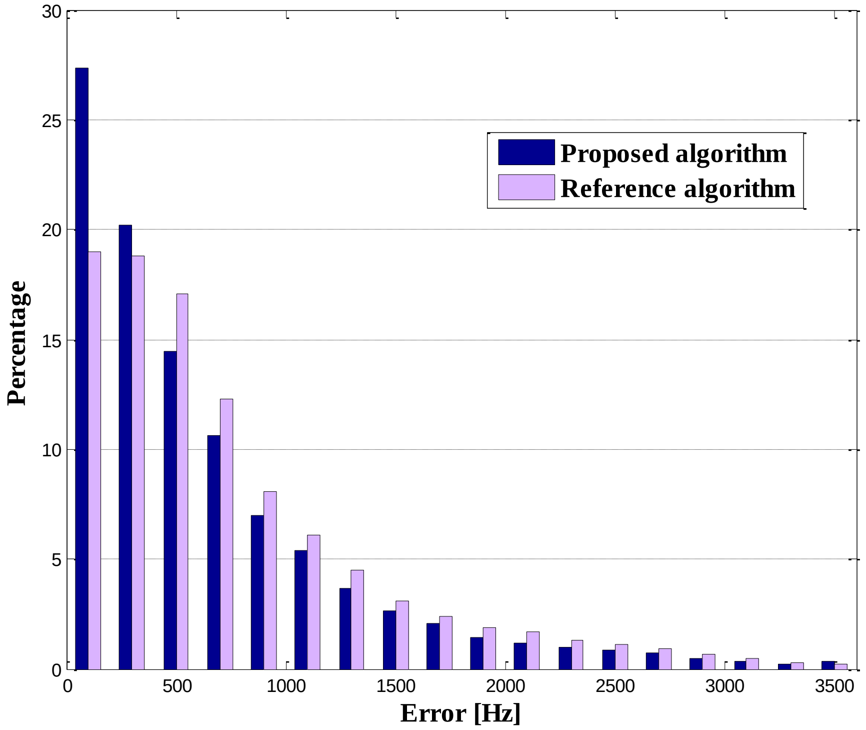


Figure 5.5: Histogram of estimated formants frequencies error.

mated formant frequencies and their original counterparts. Formants locations for the same phoneme may be different for different speakers. For comparison reason with [13] a linear predictor of order 14 was used and the formant frequencies were derived by peak picking in the spectral envelope. This measure was calculated for voiced frames in the entire test database using the TIMIT transcription.

Our results are compared in Fig. 5.5 to the results obtained in [13]. The X axis represents the formants error in Hz, the Y axis represents the percentage of estimated formants in the range. Each histogram bin has a width of 200 Hz. The results demonstrate an improvement in formant frequencies estimation using the vocal tract shape modeling and tuning.

The last objective measure criteria was the SDM. It was used to evaluate the significance of the iterative postprocessing step for quality improvement. The measure was taken between the original WB spectral envelopes and the estimated spectral envelopes with and without the iterative postprocessing step. This measure was taken only for voiced estimated frames using the HMM phoneme estimation output. The SDM parameters were set to $\alpha = 0.1$ and $\beta = 5$. The distortion was calculated using the FFT bin

Table 5.1: Average SDM and LSD of estimated spectral envelope with and without the iterative postprocessing step.

| **Measured** | **SDM [dB]** | **LSD [dB]** |
|---|---|---|
| Without iterative process | 13.6380 | 9.9759 |
| With iterative process | 9.8889 | 9.9057 |

indices from $k_{\text{low}}$ to $k_{\text{high}}$, corresponding to the frequency range from 3.4 to 8 kHz.

The mean SDM and LSD of the entire test database is presented in Table 5.1. It can be seen that the iterative postprocessing step improves the quality of the estimated spectral envelope and hence reducing the SDM. It is also noticeable from the SDM and LSD results that the achieved improvement of the iterative process is shown clearly in the SDM results and barely shown in the LSD results. This means that the postprocessing step reduces the spectral envelope over-estimation and reduces distortion in the low part of the HB frequencies which are more significant for human perception.

### 5.2.2 Subjective Evaluation

The subjective MUSHRA test was performed by 11 listeners. The test included 6 different experiments, English sentences, 3 by male and 3 by female. All experiments included a WB reference speech signal, a NB anchor speech signal, the proposed BWE speech signal and a reference BWE speech signal. The reference BWE speech signal was based on the algorithm from [20] with some unpublished improvements made by Bernd Geiser until 2010. In the test, the listeners compared multiple conditions of a sample at the same time, and could repeat the samples. The listeners could also repeat the reference when they wanted. The test produced results for the conditions between 0 and 100, with 100 being same quality as the reference speech signal.

The results of the MUSHRA test are presented in Fig. 5.6. The obtained results indicate that the proposed BWE algorithm improves the received NB signal. It also exhibits some improvement over the reference algorithm results.

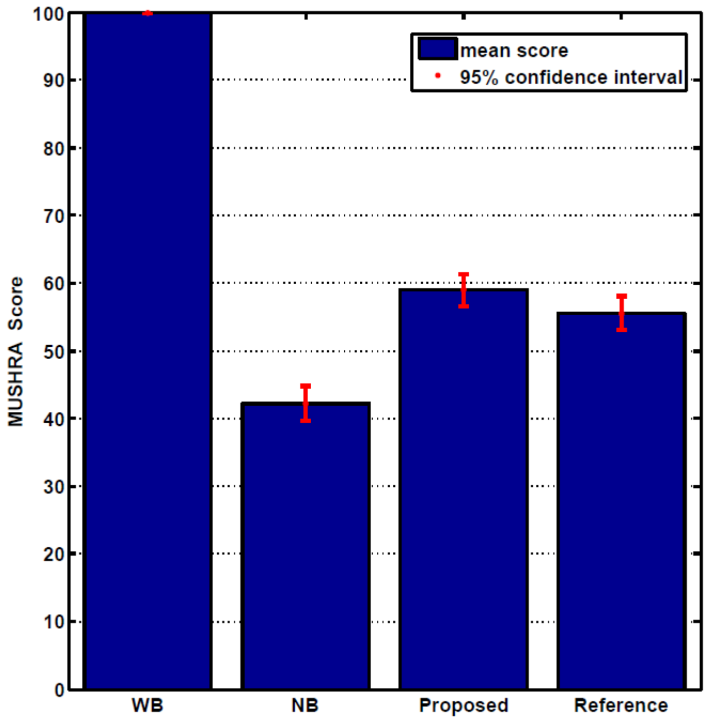


Figure 5.6: MUSHRA subjective measure score.

### 5.2.3 Complexity Evaluation

The algorithm was also examined for its complexity. The goal of this examination is to detect the most complex stages in the algorithm. This examination was performed by measuring the Matlab processing time of each major algorithm processing block, running on Intel CPU at 2.66 GHz clock speed. The result was averaged over the entire speech frames of the TIMIT test database. The distinct blocks are:

- Preprocessing and feature extraction of Stage I, as described in Section 4.1.
- State estimation of the first step of Stage II, as described in Subsection 4.2.1.
- WB VTAF estimation of the second step of Stage II, as described in Subsection 4.2.2.
- An iterative process block in the postprocessing of the third step of Stage II, as described in Subsection 4.2.3.
- Gain adjustment in the postprocessing of the third step of Stage II, as described in Subsection 4.2.3.

Table 5.2: Average computation time, using Matlab, of main BWE algorithm processing blocks.

| **Algorithm Processing Block** | **Computation Time [msec]** |
|---|---|
| Preprocessing and feature extraction | 1.27 |
| State estimation | 19.39 |
| WB VTAF estimation | 0.59 |
| Postprocessing (iterative process) | 7.69 |
| Postprocessing (gain adjustment) | 0.36 |
| WB excitation generation | 0.04 |
| WB speech synthesis | 0.57 |
| **Total** | 29.91 |

- WB excitation generation of Stage III, as described in Section 4.3.

- The WB speech synthesis of Stage IV, as described in Section 4.4.

The obtained results are presented in Table 5.2. The results indicate the average processing time of a 20 msec speech frame. The results reveal that the HMM-based state estimation step in the WB spectral envelope estimation stage consumes the most processing time. This is caused mainly because of the GMM probability values calculation. The postprocessing step also exhibit high computation load due to the on-line sensitivity function calculation, which consumes about 85% of this step processing time.

## 5.3 Summary

We have presented the performance evaluation of the proposed BWE algorithm. The algorithm was evaluated by objective and subjective measurements and was compared to other state-of-the-art BWE algorithms. The LSD measure in different phonetic categories showed improved performance of the proposed algorithm compared to the results obtained in [32]. The formant frequencies error between HB estimated formant frequencies and their original counterparts was compared to the results obtained in [13]. The results showed improved HB formant frequencies estimation using the vocal tract shape modeling and tuning. The SDM of the estimated spectral envelopes with and without the iterative postprocessing step showed the importance of the postprocessing step for improved estimation of voiced frames.

The MUSHRA subjective test was used to evaluate the proposed algorithm compared to the algorithm input NB signal and a reference BWE signal. The MUSHRA score indicates that the proposed BWE algorithm improves the received NB signal. It also exhibits some improvement over the reference algorithm results.

Complexity measure of the algorithm main building blocks showed the high complexity of the phoneme estimation step and the iterative process in the WB spectral envelope estimation stage.

# Chapter 6

# Conclusion

## 6.1 Summary

In this work, we have presented a new approach for speech bandwidth extension (BWE). Bandwidth extension is motivated by the limited frequency range in ordinary telephone networks. This limitation reduces speech quality, speech naturalness, and speech intelligibility. The aim of the BWE algorithm is to estimate the information of the missing frequency bands and improve received NB speech quality.
We have addressed two major problems of existing BWE algorithms:

- Estimation of the WB vocal tract filter - this is crucial for high intelligibility speech.
- Estimation of missing bands gain associated to the input NB gain - this is crucial for high quality of estimated WB speech without adding any artifacts.

The proposed BWE algorithm is based on the speech source-filter model. This allows the separation of the high-band signal estimation into two independent procedures, for spectral envelope estimation and for excitation generation. The algorithm involves both phoneme dependent and speaker dependent estimation of the spectral envelope. A three-step estimation algorithm was developed to deal with common difficulties in spectral envelope estimation. These difficulties are the estimation of unvoiced sounds and the robustness to different speakers and to erroneous estimation. The phoneme estimation employs an HMM to estimate the phonetic content of a speech frame. The spectral envelope estimation relies on a CB searching to estimate the speaker's VTAF. Postprocessing of the initial estimated WB VTAF, by matching formant frequencies in the low band to those of the input NB speech, smoothing in time, and gain adjustment, improved the HB spectral envelope estimation.

The proposed algorithm was fully implemented using Matlab. The experimental results demonstrate the improved performance of the proposed algorithm compared to other methods. The log spectral distance (LSD) between the estimated power spectrum and the original wideband power spectrum for different phoneme categories shows improved results compared to other BWE algorithm based on speech sounds classification. These improved results illustrate the effectiveness of phoneme dependent estimation of BWE speech frame.

The formant frequency error between the estimated highband formants and the original highband formant shows improved results compared to other BWE algorithm based on the acoustic tube model. These improved results illustrate the effectiveness of estimating the speaker VTAF using the codebook search.

The spectral distortion measure (SDM) between the estimated highband spectral envelope and the original highband spectral envelope shows improved results of about 4 dB for frames which were postprocessed using the iterative tuning compared to those that weren't. These results illustrate the effectiveness of the iterative tuning process for estimation artifacts reduction.

The MUSHRA listening tests show improved quality of the enhanced speech. An improvement of more than 15 points compared to the input narrowband speech illustrates the advantage of using the proposed BWE algorithm when using old telephone networks. The proposed BWE algorithm could be used to improve the call quality in the transition time to real wideband networks.

The drawbacks of the proposed algorithm are twofold. First, the concatenated tube model is limited in modeling VTAF shape of unvoiced and nasal sounds. Second, the HMM based phoneme estimation and the online sensitivity function calculation at the postprocessing step, require high computational complexity.

## 6.2 Future Research

To deal with the proposed BWE algorithm drawbacks, further research needs to be carried out. Future work might include a different VTAF estimation technique for unvoiced and nasal sounds. Using the postprocessing iterative procedure for better refinement and control of estimated spectral envelope by high-band formants tuning to past estimated high-band formants could improve the smoothing in time of the estimated HB spectral envelope. This should allow improved estimated speech signal quality. Offline calculation of the sensitivity function, for each WB VTAF codeword, will reduce the computational complexity. Using normalized formant frequency deviation for the iterative process, might reduce the needed number of iterations, while maintaining the same achieved postprocessing quality. The algorithm should also be evaluated using formal listening tests, under different background noise conditions and with different languages. This evaluation would determine the algorithm robustness to noisy environments and multiple languages.

Bandwidth extension of speech in this research was focused on improving old telephone networks speech quality. Even after a transition phase to wideband supported telephone networks, there will be a need for BWE. It can be incorporated into wideband speech codecs, in order to maintain low bit rates. The concept of BWE can easily be extended to estimating "super-wideband", i.e. extending speech signals bandwidth from 8 kHz to 16 kHz. It can also be used in noise cancellation algorithm for estimation and synthesis of clean speech in frequency bands with low signal to noise ratio (SNR). Finally historical recordings could be bandwidth extended in order to obtain a wideband version of the recordings.

# Appendix A

# Acoustic Tube Model of the Vocal Tract

In this appendix, we describe the acoustic tube model of the vocal tract. Chapter 2, presented a method to model the vocal tract as an AR filter. Another possibility of modeling the vocal tract comprises the use of a tube. The vocal tract is regarded as an acoustic tube with a varying cross-sectional area. The tube is divided into an arbitrary number, $N_A$, of sections with equal length $l$. The following assumptions are made for this model.

- *Assumption 1*: The transverse dimension of each section is small enough compared with a wavelength so that the wave propagating through one section can be treated as a plane wave.
- *Assumption 2*: The tube is rigid, and the losses in the sound wave due to viscosity and heat conduction are negligible.

The Webster's horn equation represents a plain wave that propagates through a lossless cylindrical tube with a variable cross-sectional area $A(x)$.

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{1}{A}\frac{dA}{dx}\frac{\partial \Phi}{\partial x} = \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} \quad , \tag{A.1}$$

where $c$ represents the sound propagation velocity, $x$ the longitudinal axis along the tube, $t$ the time, and $\Phi(x,t)$ the so-called velocity potential. The velocity potential is defined

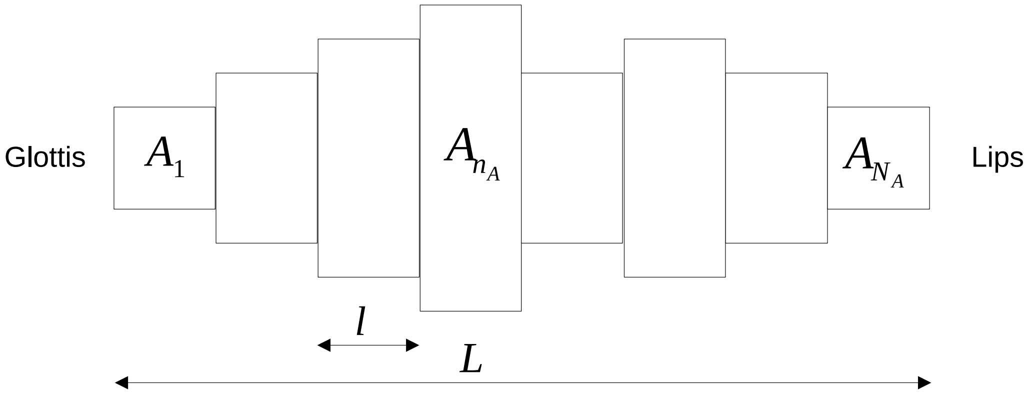


Figure A.1: Model of the vocal tract consisting of a concatenation of several lossless cylindrical tubes of length $l$ and cross-sectional area $A_{n_A}$.

by:

$$\begin{aligned} p &= \rho_0 \tfrac{\partial \Phi}{\partial t} \quad \text{and} \\ v &= -\tfrac{\partial \Phi}{\partial x} \quad , \end{aligned} \tag{A.2}$$

where $\rho_0$ denotes the density of air in the tube, $p$ the pressure, and $v$ the sound particle velocity. As a rule, closed-form solutions for this problem are not available.

If we assume a model corresponding to Fig. A.1 we can reformulate (A.1) for each tube segment with a constant cross-sectional area $A_{n_A}$ to

$$\frac{\partial^2 \Phi_{n_A}}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \Phi_{n_A}}{\partial t^2} \quad , \tag{A.3}$$

A general solution to (A.3) is of the form

$$\Phi_{n_A}(x,t) = \Phi^+_{n_A}\left(t - \frac{x}{c}\right) + \Phi^-_{n_A}\left(t + \frac{x}{c}\right) \quad . \tag{A.4}$$

Here $\Phi^+_{n_A}$ denotes the wave that propagates in positive x-direction (forward) of segment $n_A$, whereas $\Phi^-_{n_A}$ denotes the wave that propagates in negative x-direction (backward) within the tube segment $n_A$, where x is measured from the glottal end of each tube $(0 \leq x \leq l)$.

In the case of one-dimensional airflow in the vocal tract, it is more convenient to look at the velocity of a volume of air $u_{n_A}$ than its particle velocity $v_{n_A}$. Let define the volume velocity as $u_{n_A} = A_{n_A} v_{n_A}$ and the acoustic impedance as $Z_{n_A} = \frac{\rho_0 c}{A_{n_A}}$. Using (A.4) in (A.2) we can formulate the volume velocity and the pressure as

$$\begin{aligned} u_{n_A}(x,t) &= \tfrac{A_{n_A}}{c}\left[\Phi^+_{n_A}\left(t - \tfrac{x}{c}\right) - \Phi^-_{n_A}\left(t + \tfrac{x}{c}\right)\right] = \\ &= u^+_{n_A}\left(t - \tfrac{x}{c}\right) - u^-_{n_A}\left(t + \tfrac{x}{c}\right) \quad , \end{aligned} \tag{A.5}$$

and

$$\begin{aligned} p_{n_A}(x,t) &= \rho_0 \left[\Phi^+_{n_A}\left(t - \tfrac{x}{c}\right) + \Phi^-_{n_A}\left(t + \tfrac{x}{c}\right)\right] = \\ &= p^+_{n_A}\left(t - \tfrac{x}{c}\right) + p^-_{n_A}\left(t + \tfrac{x}{c}\right) = \\ &= Z_{n_A}\left[p^+_{n_A}\left(t - \tfrac{x}{c}\right) + p^-_{n_A}\left(t + \tfrac{x}{c}\right)\right] \quad . \end{aligned} \tag{A.6}$$

Now the volume velocity and the sound pressure must be continuous at the junction between section $n_A$ and the subsequent one $n_A + 1$, resulting in

$$\begin{aligned} u_{n_A}(l,t) &= u_{n_A+1}(0,t) \\ p_{n_A}(l,t) &= p_{n_A+1}(0,t) \quad . \end{aligned} \tag{A.7}$$

We can express the volume velocity and sound pressure for the first case as

$$u_{n_A}(l,t) = u^+_{n_A}(t-\tau) - u^-_{n_A}(t+\tau) \tag{A.8}$$

$$p_{n_A}(l,t) = Z_{n_A}\left[u^+_{n_A}(t-\tau) + u^-_{n_A}(t+\tau)\right] \tag{A.9}$$

where $\tau = \frac{l}{c}$ denotes the time needed by the wave to propagate through the tube segment $n_A$ with length $l$. In the same manner we can express the volume velocity and the sound pressure for the second case as

$$u_{n_A+1}(0,t) = u^+_{n_A+1}(t) - u^-_{n_A+1}(t) \tag{A.10}$$

$$p_{n_A+1}(0,t) = Z_{n_A+1}\left[u^+_{n_A+1}(t) + u^-_{n_A+1}(t)\right] \tag{A.11}$$

Inserting both conditions into (A.7) leads to

$$u^+_{n_A}(t-\tau) - u^-_{n_A}(t+\tau) = u^+_{n_A+1}(t) - u^-_{n_A+1}(t) \tag{A.12}$$

$$u^+_{n_A}(t-\tau) + u^-_{n_A}(t+\tau) = \frac{A_{n_A}}{A_{n_A+1}}\left[u^+_{n_A+1}(t) + u^-_{n_A+1}(t)\right] \quad . \tag{A.13}$$

The propagation of the volume velocity between two subsequent tube section is presented in Fig. A.2. If we now assume the case of $u^-_{n_A+1}(t) = 0$ then $u^-_{n_A}(t+\tau)$ can be interpreted as the reflection of $u^+_{n_A}(t-\tau)$ at the transition of tube segment $n_A$ to $n_A + 1$. Setting $u^-_{n_A+1}(t) = 0$ in (A.12) and (A.13) yields

$$u^-_{n_A}(t+\tau) = \frac{A_{n_A} - A_{n_A+1}}{A_{n_A} + A_{n_A+1}} u^+_{n_A}(t-\tau) \tag{A.14}$$

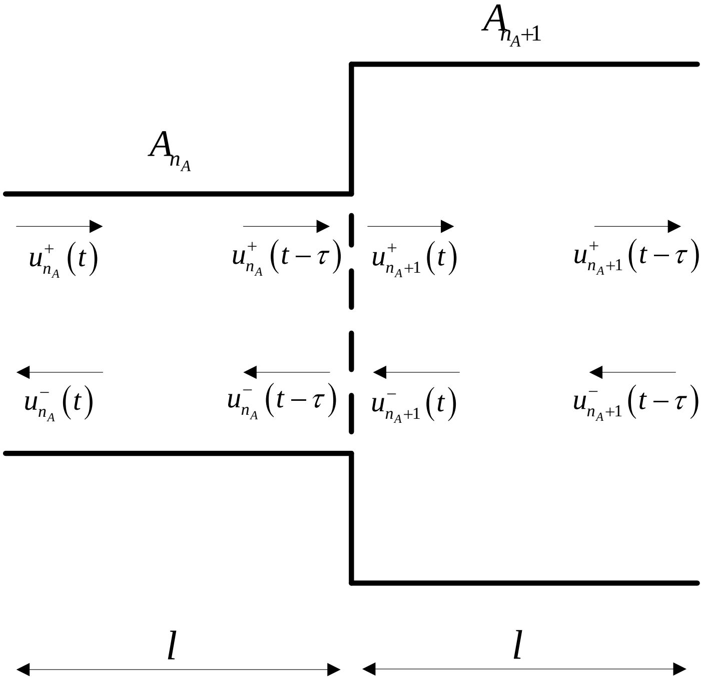


Figure A.2: Volume velocity propagation in the acoustic tube.

Thus we can define the reflection coefficient as

$$R_{n_A} = \frac{u^-_{n_A}(t+\tau)}{u^+_{n_A}(t-\tau)} = \frac{A_{n_A} - A_{n_A+1}}{A_{n_A} + A_{n_A+1}} \tag{A.15}$$

The filtering processes of the linear prediction model and an acoustic tube model of speech are equivalent [42] , with the reflection coefficients in the acoustic tube model as a common factor. Now, from the reflection coefficients that can be derived as a byproduct of the Levinson-Durbin algorithm of LP analysis, we can derive the VTAF by

$$A_{n_A} = \frac{1+R_{n_A}}{1-R_{n_A}} A_{n_A+1} \tag{A.16}$$

# Appendix B

# Equalizer Lowpass Filter Characteristics

In this appendix, we present the characteristics of the equalizer lowpass filter, described in Chapter 4. The equalizer lowpass filter is a stable, real discrete time FIR filter with the following characteristics:

Filter structure: direct-form II transposed.

Numerator length: 80.

Numerator coefficients:

0.0019 0.0013 0.0001 -0.0016 -0.002 -0.0021 -0.0044 -0.007 -0.0066 -0.0056 -0.0081 -0.0108 -0.0089 -0.0064 -0.0081 -0.0093 -0.0059 -0.0024 -0.0016 0.0005 0.0044 0.0072 0.0123 0.0199 0.0217 0.0206 0.0322 0.0482 0.0433 0.0325 0.0542 0.084 0.0646 0.0325 0.0732 0.1341 0.0801 -0.0196 0.0842 0.4023 0.5857 0.4023 0.0842 -0.0197 0.08 0.134 0.0733 0.0324 0.0644 0.0839 0.0543 0.0325 0.0432 0.0482 0.0322 0.0206 0.0216 0.0199 0.0123 0.0072 0.0044 0.0005 - 0.0017 -0.0024 -0.0058 -0.0093 -0.0082 -0.0064 -0.0088 -0.0108 -0.0082 -0.0056 -0.0065 -0.007 -0.0044 -0.0021 -0.0021 -0.0016 0.0001 0.0013

Denominator length: 1.

Denominator coefficient:

1

The filter frequency response is presented in Fig B.1.

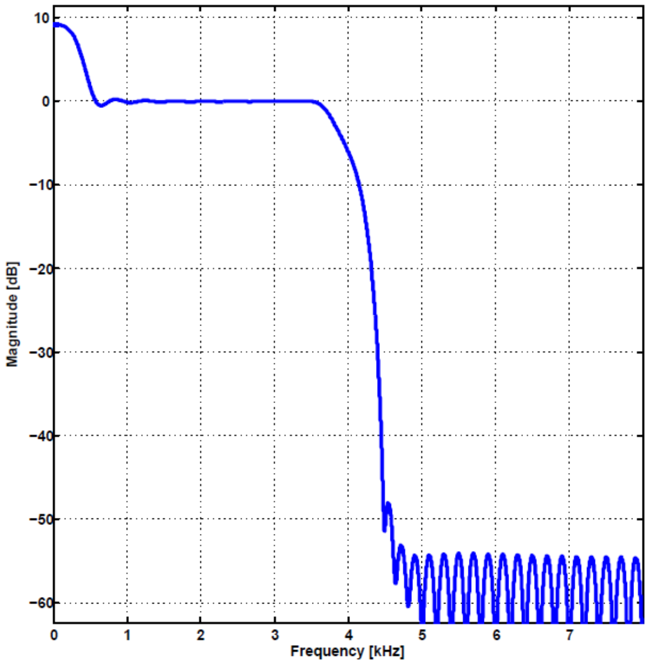


Figure B.1: Frequency response of the equalizer lowpass filter.

# Appendix C

# Definition of Mutual Information and Separability

In this appendix, the mutual information and separability measures from information theory and statistics will be reviewed, these measures are described in Chapter 2.

Shannon's mutual information $I(\mathbf{x} : \mathbf{y})$ yields the mean information we gain on the estimated WB spectral envelope representation $\mathbf{y}$ by knowledge of the feature vector $\mathbf{x}$. Mutual information is an indication of the feasibility and quality of the estimation task [19, 29, 30]. Let $E\left\{\|\mathbf{y}-\tilde{\mathbf{y}}\|^2\right\}$ denote the mean-square estimation error between the real WB spectral envelope representation $\mathbf{y}$, and the estimated WB envelope representation $\tilde{\mathbf{y}}$. It has been shown in [17], that the minimum achievable mean-squared estimation error is lower bounded by a value dependent on the mutual information. The larger the mutual information, the lower is the bound. Hence, a large mutual information $I(\mathbf{x} : \mathbf{y})$ is a necessary condition for high-quality estimation of $\mathbf{y}$ from the observations $\mathbf{x}$.

From the field of pattern recognition, the separability is known as a measure of the quality of a particular feature set for a classification problem [12]. In the BWE algorithm, the class definitions is adopted in the method used to estimate the WB spectral envelope. I.e., if codebook mapping is used, the classes should correspond to the correct codebook indices as computed from true WB speech. For an HMM-based approach, the classes should be the true HMM state information.

The separability measure can be calculated from a tagged set of training data, that is, for each feature vector in the set the corresponding class must be known (i.e., the

WB corresponding mapping). Let $\mathbf{X_i}$ denote the set of feature vectors $\mathbf{x}$ assigned to the $i$th class. The number of feature vectors in the $i$th set is $N_{\mathbf{X_i}} = |\mathbf{X_i}|$. The constant $N_S$ denotes the number of classes. Then, the total number of frames in the training data is given by $N_m = \sum_{i=1}^{N_S} N_{\mathbf{X_i}}$. From the tagged training data, the within-class covariance matrix is calculated by

$$\mathbf{V_x} = \frac{1}{N_m} \sum_{i=1}^{N_S} \sum_{\mathbf{x} \in \mathbf{X_i}} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T \ . \tag{C.1}$$

The between-class covariance matrix is calculated by

$$\mathbf{B_x} = \sum_{i=1}^{N_S} \frac{N_{\mathbf{X_i}}}{N_m} (\mu_i - \mu)(\mu_i - \mu)^T \ , \tag{C.2}$$

where

$$\mu_i = \frac{1}{N_{\mathbf{X_i}}} \sum_{\mathbf{x} \in \mathbf{X_i}} \mathbf{x}, \quad \mu = \sum_{i=1}^{N_S} \frac{N_{\mathbf{X_i}}}{N_m} \mu_i \ . \tag{C.3}$$

The separability measure will be larger if the between-class covariance gets larger and/or if the within-class covariance gets smaller. Therefore, the separability measure is empirically defined by the term $\mathbf{J_x} = (\mathbf{V_x^{-1} B_x})$. To obtain a scalar measure for the separability of the classes, a trace criterion is used [12] to obtain the separability measure $\zeta(\mathbf{x})$

$$\zeta(\mathbf{x}) = \mathrm{tr}(\mathbf{J_x}) = \mathrm{tr}\left(\mathbf{V_x^{-1} B_x}\right) \tag{C.4}$$

The separability depends on the definition of the classes. Comparing $\zeta(\mathbf{x})$ for different feature vectors $\mathbf{x}$ with the same class definitions, a larger value indicates a better suitability of the corresponding feature vector for classification and estimation. The separability measure has the following properties:

- The definition of the separability measure is based on the implicit assumption of a normal distribution of the feature vectors that are assigned to each class. If this assumption is not valid, the significance of the separability measure is reduced.

- By the separability measure, all classes are treated alike. Therefore, the separability of two very similar classes (w.r.t. the represented speech sound) is rated like the separability of two very different classes. Hence, maximizing the separability does not necessarily lead to the optimum achievable estimation performance (e.g., in the MMSE sense) of the subsequent estimation rule.

- In general, the values of the separability cannot be added up if several features are assembled to a composite feature vector. In this case, the separability of the composite feature vector must be measured again.

# Appendix D

# MUSHRA Evaluation Listening Test Results

In this appendix, we present the test setup and obtained results of the MUSHRA test, described in Chapter 5. The MUSHRA tests were conducted in a quiet room. The speech files were played on a standard PC using a set of Sony MDR-XD100 headphones. Before the test, a short introduction was given and the test listener was then told to read a short presentation guide for the MUSHRA listening test. The average test duration was about 15 minutes. Table D presents the obtained listening tests results.

Table D.1: MUSHRA listening tests results.

| **Listener 1** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 39 | 38 | 28 | 44 | 39 | 25 |
| Proposed | 82 | 86 | 80 | 70 | 83 | 78 |
| Reference | 69 | 70 | 61 | 52 | 64 | 72 |
| **Listener 2** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 60 | 61 | 50 | 68 | 78 | 85 |
| Proposed | 50 | 71 | 76 | 55 | 76 | 85 |
| Reference | 70 | 69 | 85 | 74 | 73 | 85 |
| **Listener 3** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 19 | 50 | 30 | 30 | 30 | 50 |
| Proposed | 70 | 70 | 50 | 50 | 70 | 50 |
| Reference | 70 | 70 | 30 | 30 | 50 | 71 |
| **Listener 4** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 50 | 33 | 50 | 50 | 50 | 50 |
| Proposed | 27 | 50 | 31 | 60 | 50 | 50 |
| Reference | 27 | 43 | 36 | 50 | 41 | 50 |
| **Listener 5** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 40 | 13 | 27 | 50 | 30 | 30 |
| Proposed | 80 | 57 | 70 | 24 | 50 | 67 |
| Reference | 50 | 75 | 61 | 75 | 73 | 50 |
| **Listener 6** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 0 | 0 | 30 | 50 | 10 | 70 |
| Proposed | 30 | 40 | 50 | 13 | 27 | 50 |
| Reference | 70 | 40 | 20 | 50 | 50 | 80 |

| **Listener 7** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 0 | 18 | 0 | 15 | 15 | 15 |
| Proposed | 70 | 0 | 69 | 66 | 68 | 46 |
| Reference | 36 | 32 | 31 | 30 | 22 | 50 |

| **Listener 8** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 54 | 53 | 59 | 58 | 55 | 57 |
| Proposed | 68 | 62 | 61 | 68 | 66 | 86 |
| Reference | 85 | 80 | 68 | 82 | 82 | 65 |

| **Listener 9** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 50 | 50 | 70 | 60 | 60 | 60 |
| Proposed | 80 | 80 | 80 | 50 | 70 | 70 |
| Reference | 40 | 70 | 60 | 40 | 40 | 70 |

| **Listener 10** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 10 | 70 | 30 | 30 | 30 | 50 |
| Proposed | 30 | 50 | 50 | 70 | 90 | 70 |
| Reference | 90 | 80 | 10 | 50 | 70 | 90 |

| **Listener 11** | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 5 | Trial 6 |
|---|---|---|---|---|---|---|
| WB | 100 | 100 | 100 | 100 | 100 | 100 |
| NB | 50 | 81 | 70 | 50 | 56 | 50 |
| Proposed | 90 | 81 | 27 | 33 | 23 | 41 |
| Reference | 80 | 18 | 19 | 28 | 9 | 32 |

# Bibliography

[1] International Telecommunications Union, ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA). ITU Publications: http://www.itu.int/publications/default.aspx?menu=main

[2] International Telecommunications Union, ITU-T Recommendation G.722, "7 kHz audio-coding within 64 kbit/s," November 1988, Available: http://www.itu.int.

[3] International Telecommunications Union, ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," ITU 2001, Available: http://www.itu.int.

[4] International Telecommunications Union, ITU-T Recommendation P.48 "Specification for an intermediate reference system" ITU-T publications, Nov. 1988, Available: http://www.itu.int.

[5] B. S. Atal, "Determination of the vocal tract shape directly from the speech wave," *Journal of the Acoustical Society of America*, vol. 47, p. 65, 1970.

[6] P. Bauer, and T. Fingscheidt, "A statistical framework for artificial bandwidth extension exploiting speech waveform and phonetic transcription," in *Proc. European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, August 2009, pp. 1839–1843.

[7] D. Chow and W. H. Abdulla, "Robust speaker identification based on perceptual log area ratio and Gaussian mixture models," in *Proc. INTERSPEECH 2004 ICSLP*, Jeju Island, Korea, October 2004, p. 1761–1764.

[8] R. V. Cox, D. Malah and D. Kapilov, "Improving upon toll quality speech for VoIP," in *Proc. 38'th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, USA, Nov. 7-10, 2004, pp. 405–409.

[9] J. R. Deller Jr., J. H. L. Hansen and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* IEEE Press, Piscataway, NJ, USA, 2000.

[10] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the Mel frequency cepstral coefficients," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999, pp. 171–173.

[11] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999, pp. 174–176.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann, Academic Press, San Francisco, San Diego, 2nd edition, 1990.

[13] H. Gustafsson, U. A. Lindgren and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 577–588, 2006.

[14] J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

[15] B. Iser, W. Minker and G. Schmidt, *Bandwidth extension of speech signals.* Lecture Notes in Electrical Engineering, vol. 13, Springer, 2008.

[16] L. H. Jamieson. "Introduction to the physiology of speech and hearing." Available: http://cobweb.ecn.purdue.edu/ ee649/notes/physiology.html.

[17] P. Jax, "Bandwidth Extension for Speech," Chapter 6, *Audio bandwidth extension*, Larsen and Aarts, Eds., Wiley, November 2004.

[18] P. Jax and P. Vary, "Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?," *IEEE communication magazine*, vol. 44, no. 5, pp. 106–111, May 2006.

[19] P. Jax and P. Vary, “Feature selection for improved bandwidth extension of speech signals,” in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process. (ICASSP 2004)*, Montreal, Quebec, Canada, May 2004, pp. 697–700.

[20] P. Jax and P. Vary, “On artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, August 2003.

[21] U. Kornagel, “Improved Artificial Low-Pass Extension of Telephone Speech,” *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, Kyoto, Japan, Sept. 2003, pp. 107–110.

[22] U. Kornagel, “Techniques for artificial bandwidth extension of telephone speech,” *Signal Processing*, vol. 86, pp. 1296–1306, 2006.

[23] S. Krstulovi’c, “Acoustico-articulatory inversion of unequal-length tube models through lattice inverse filtering,”, IDIAP, Tech. Rep. Idiap-RR-16-1998, 1998.

[24] S. Krstulovi’c, “Speech Analysis with Production Constraints,” PhD thesis, Ecole Polytechnique Federale de Lausanne, 2001.

[25] D. Malah, “Method of bandwidth extension for narrow-band speech,” US Patent 6988066 B2, Jan 2006.

[26] P. Mokhtari and F. Clermont, “New perspectives on linear-prediction modelling of the vocal-tract: uniqueness, formant-dependence and shape parameterisation,” *Proc. 8th Australian Int. Conf. on Speech Science and Tech*, Canberra, Australia, 2000, pp. 478-483.

[27] T.K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process.* Mag. 13 (6), pp. 47–60, November 1996.

[28] M. Nilsson and W. B. Kleijn, “Avoiding over-estimation in bandwidth extension of telephony speech,” in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process. (ICASSP 2001)*, Salt Lake City, UT, USA, May 2001, pp. 869–872.

[29] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, “Gaussian mixture model based mutual information estimation between frequency bands in speech,” in

*Proc. IEEE Int. Conf. Acoust, Speech, Signal Process. (ICASSP 2002)*, Orlando, FL, USA, May 2002, pp. 525–528.

[30] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech," in *Proc. INTERSPEECH 2008*, Brisbane, Australia, Sept. 2008, pp. 53–56.

[31] H. Pulakka, V. Myllyla, L. Laaksonen, P. Alku, "Bandwidth extension of telephone speech using a filter bank implementation for highband Mel spectrum," in *Proc. European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, August 2010, pp. 979–983.

[32] H. Pulakka, L. Laaksonen, M. Vainio, J. Pohjalainen and P. Alku, "Evaluation of an artificial speech bandwidth extension method in three languages," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 6, pp. 1124–1137, 2008.

[33] Y. Qian and P. Kabal, "Dual-Mode Wideband Speech Recovery from Narrowband Speech", in *Proc. 8th European Conf. Speech, Commun. Tech.*, Geneva, Italy, Sept. 2003, pp. 1433–1437.

[34] S. Quackenbush, T. Barnwell, and Clements, *Objective measures of speech quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[35] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[36] T. Ramabadran and M. Jasiuk, "Artificial bandwidth extension of narrow-band speech signals via high-band energy estimation," in *Proc. European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, August 25-29. 2008.

[37] J. Schoentgen "Speech Modelling Based on Acoustic-to-Articulatory Mapping," Summer School on Neural Networks, volume 3445, 2004.

[38] G. Song and P. Martynovich, "A study of HMM-based bandwidth extension of speech signals," *Signal Processing*, vol. 89, pp. 2036–2044, October 2009.

[39] J. L. Stevens, *Linear algebra with applications*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[40] B. Story, "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions," *J. Acoust. Soc. Amer*, vol. 119, pp. 715–718, 2006.

[41] L.M. Surhone, M.T. Surhone and S.F. Henssonow *MUSHRA*. VDM Verlag Dr. Mueller AG & Co. Kg, 2010.

[42] H .Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. on Audio and Electroac.*, vol. 21, pp. 417–427, 1973.

[43] H .Wakita, "Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 281-285, June 1979.

[44] H. Yasukawa, "Quality enhancement of band limited speech by filtering and multirate techniques", *Proc. ICSLP 1994*, September. 1994, pp. 1607–1610.

[45] H. Yasukawa. "Restoration of wide band signal from telephone speech using linear prediction error processing", in *Proc. ICSLP 96*, Philadelphia, PA, USA, Oct. 1996, vol. 2, pp. 901–904.

[46] H. Yasukawa. "Signal restoration of broad band speech using nonlinear processing", in *Proc. European Signal Processing Conference (EUSIPCO 1996)*, Trieste, Italy, Sept. 1996, pp 987–990.