HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Laura Kallio

Artificial Bandwidth Expansion of Narrowband Speech in Mobile Communication Systems

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, Dec 9, 2002

Supervisor: Professor Paavo Alku

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Laura Kallio
Name of the thesis:	Artificial Bandwidth Expansion of Narrowband Speech in Mobile Communication Systems
Date:	Dec 9, 2002 **Number of pages:** 53+9
Department:	Electrical and Communications Engineering
Professorship:	S-89
Supervisor:	Prof. Paavo Alku

In most of the communication systems speech is transmitted in narrowband, containing frequencies from 300 Hz to 3400 Hz. The bandwidth of a narrowband speech signal can be expanded using a method called artificial bandwidth expansion. The idea is that by increasing the bandwidth and by creating artificially new frequency components to the signal, it is possible to enhance the speech quality with existing communication systems.

In this thesis different frequency expansion methods are discussed. They are based on various approaches to the problem. In addition different ways to evaluate speech quality have been studied. They can be divided into subjective and objective methods.

Within this thesis an artificial bandwidth expansion algorithm for telephone band speech has been developed further. The algorithm is based on aliasing which increases the sampling rate from 8 kHz to 16 kHz and creates a mirror image of the narrowband spectrum to the highband. Spectra of different speech sounds are modified differently in order to create speech that sounds natural. In this Thesis the main objective was to improve the detection and processing of fricatives and stop-consonants.

After the development of the algorithm the quality of it was evaluated through subjective listening tests. The test was so called paired comparison test. The purpose was to find out whether people prefer the old narrowband telephone speech or artificially expanded speech. Different processing cases simulating possible options for how the bandwidth expansion could be implemented in telecommunication systems were included in the test. The results of the listening test show that the algorithm is working but some improvements are still needed.

Keywords: speech processing, artificial bandwidth expansion, aliasing, speech quality, subjective listening test

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Laura Kallio	
Työn nimi:	Kapeakaistaisen puheen taajuuskaistan keinotekoinen laajentaminen matkapuhelinympäristössä	
Päivämäärä:	9.12.2002	**Sivuja:** 53+9
Osasto:	Sähkö- ja tietoliikennetekniikka	
Professuuri:	S-89	
Työn valvoja:	Prof. Paavo Alku	

Puhetta siirretään tiedonsiirtojärjestelmissä yleensä kapeakaistaisena, eli signaali sisältää taajuuksia väliltä 300-3400 Hz. Kapeakaistaisen puhesignaalin taajuuskaistaa voidaan laajentaa keinotekoisella kaistanlaajennusmenetelmällä. Ajatuksena on, että kasvattamalla taajuuskaistaa ja luomalla keinotekoisesti puhesignaaliin uusia taajuuskomponentteja, voidaan puheen laatua parantaa käyttämällä olemassa olevia tiedonsiirtojärjestelmiä.

Tässä työssä on esitelty erilaisia taajuuskaistan laajennusmenetelmiä. Ne eroavat toisistaan lähestymistavoiltaan. Lisäksi on tutkittu puheen laadun arvioimismenetelmiä, jotka voidaan jakaa subjektiivisiin ja objektiivisiin menetelmiin.

Tämän työn yhteydessä on kehitetty eteenpäin puhelinkaistaiselle puhesignaalille suunniteltua kaistanlaajennusalgoritmia. Algoritmi perustuu laskostamiseen, jonka avulla näyttenottotaajuutta kasvatetaan 8 kHz:istä 16 kHz:iin ja kapeakaistaisesta spektristä luodaan peilikuva yläkaistalle. Spektriä muokataan adaptiivisesti ottamalla huomioon millainen äänne on kulloinkin kyseessä. Tässä työssä tavoitteena oli parantaa erityisesti frikatiivien ja stop-konsonanttien tunnistamista ja prosessointia.

Algoritmin kehityksen jälkeen sen toimivuutta on arvioitu subjektiivisen kuuntelukokeen avulla. Toteutettu koe oli ns. parivertailukoe, jossa oli tarkoitus selvittää pitävätkö ihmiset enemmän tutusta kapeakaistaisesta “puhelinäänestä” vai keinotekoisesti laajennetusta puheesta. Kuuntelukoe sisälsi erilaisia prosessointitapauksia, joilla simuloitiin eri vaihtoehtoja miten kaistanlaajennus voitaisiin toteuttaa eri puhelinympäristöissä. Kuuntelukokeen tulokset osoittivat, että algoritmi on toimiva, mutta pienet parannukset ovat edelleen tarpeellisia.

Avainsanat: puheenkäsittely, keinotekoinen taajuuskaistan laajennus, laskostaminen, puheen laatu, subjektiivinen kuuntelukoe

Acknowledgements

This Thesis was carried out in the Laboratory of Acoustics and Audio Signal Processing of Helsinki University of Technology. The Thesis was written within a collaboration project with the Audio and Speech Processing Laboratory of Nokia Reserch Center.

I would like to express special thanks to Professor Paavo Alku for providing an interesting subject and for his valuable guidance during the project. I would also like to thank Matti Kajala, Päivi Valve and Jari Sjöberg from Nokia Research Center for the cooperation.

Otaniemi, December 9, 2002

Laura Kallio

Contents

Abbreviations

$a(k)$	Autoregressive coefficients
f_c	Cutoff frequency
f_s	Sampling frequency
f_{sb}	Stopband edge
$I(\mathbf{x};\mathbf{y})$	Mutual information
ABE	Artificial Bandwidth Expansion
ACR	Absolute Category Rating
AMR-WB	Adaptive Multi Rate - Wideband
AR	Auto-regressive
CCR	Comparison Category Rating
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
EFR	Enhanced Full Rate
EM	Expectation Maximization
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GMM	Gaussian Mixture Model
GSM	Global System for Mobile Communications
GSM1	Global System for Mobile Communications 1
HMM	Hidden Markov Model
HUT	Helsinki University of Technology
IFFT	Inverse Fast Fourier Transform
IP	Internet Protocol
IRS	Intermediate Reference System
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union

ITU-R	International Telecommunication Union - Radiocommunication Sector
ITU-T	International Telecommunication Union - Telecommunications Standardization Sector
LBG	Linde Buzo Gray -algorithm
LPC	Linear Predictive Coding
LSD	Log Spectral Distortion
LSP	Line Spectrum Pairs
MFCC	Mel-frequency Cepstral Coefficient
MOS	Mean Opinion Score
$\mathrm{MOS_{LE}}$	Mean Opinion Score - Listening Effort
$\mathrm{MOS_{Lp}}$	Mean Opinion Score - Loudness Preference
NB	Narrowband
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PSTN	Public Switched Telephone Network
RMS	Root Mean-square
SNR	Signal-to-Noise Ratio
TIMIT	Texas Instruments and the Massachusetts Institute of Technology
VAD	Voice Activity Detector
VoIP	Voice over IP
VQ	Vector Quantization
WB	Wideband

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation for Artificial Bandwidth Expansion

In most of the telecommunication systems like, for example, PSTN and GSM networks, speech is sampled with sampling frequency $f_s = 8kHz$. The Nyqvist theorem justifies that the highest preserved frequency component is half of the sampling frequency, so in this case 4 kHz. In practise, partly for historical reasons, the actual frequency band of telephone speech is even less, only 300-3400 Hz and speech of this bandwidth is here referred to as narrowband speech. Human speech contains considerably more frequencies and the lack of them can be heard in so called *telephone speech*. It is not natural but sounds muffled.

The narrowband speech signal contains enough information for traditional telephone use because the 3.1 kHz bandwidth preserves intelligibility of speech in some degree. But new applications have created new demands for the quality of speech in communication systems. For example in hands-free speaking and teleconferencing applications the narrowband speech quality is not always sufficient. In particular the separation of stop-consonants can be difficult because important higher frequencies are missing. Moreover, human preference for improving communication possibilities speaks for the need for better subjective quality of speech.

Since the oldest network, public switched telephone network (PSTN), mainly all the communication systems have transmitted speech in narrowband. The next telecommunication generation has made some promises for wider bandwidth. For instance there is a new standard for wideband speech codec, Adaptive Multi-Rate Wideband (AMR-WB) for which, according to ITU-T Recommendation G.722.2, the foreseen applications are:

– Voice over IP (VoIP) and Internet applications

- Mobile communications
- PSTN applications
- ISDN wideband telephony
- ISDN videotelephony and video-conferencing

All of these applications require good quality speech, which is why they are considered possible applications for wideband speech transmission.

Because of economic reasons it will take time before all the equipment, protocols and the whole transmission link support wideband transmission. Meanwhile another approach towards wideband transmission is to artificially add new spectral components to speech utilizing only the narrowband speech information, in for example, the receiving end of a transmission chain. In other words speech would be transmitted as before and the conversion from narrowband to wideband would be done after transmission. In practise it would usually mean that the sampling frequency would be expanded from 8 kHz to 16 kHz and missing frequency components up to 8 kHz would be constructed. Sometimes also low frequencies from 0 to 300 Hz could be added to the signal. Through this procedure which will be called **artificial bandwidth expansion** (ABE) it would be possible to obtain speech of better quality with existing communication systems. Moreover, during the change from narrowband to real wideband transmission, both narrow- and wideband signals will be transmitted in the same network and the change could be alleviated by a method that would recover the missing frequency region from narrowband signal.

The motivation for artificial bandwidth expansion methods is the fact that the narrowband signal and the missing highband signal are created by the same speech production process. Therefore, there is a reason to believe that the spectral envelope of the lower and the higher frequency bands of the speech signal are dependent and it would be possible to artificially create the highband when the narrowband is known. The basics of the human speech production system are known considerably well, but because of the versatility of aspects affecting it, a more in-depth understanding is still not available. Moreover, every human being is an individual and physical and psychological properties vary substantially. That is why the task to add new spectral components to speech using only the narrowband information is very challenging.

Because the main goal of artificial frequency expansion algorithms is to improve the quality of speech, there has to be a way to measure and evaluate quality. For this purpose there are standardized quality assessment methods that have been created to alleviate the comparison of quality features. Subjective methods measure the perceived subjective quality. They are

usually different kinds of listening tests where listeners give their opinions on sounds they hear. Objective methods are quality assessment techniques that are designed to measure some feature of a sound. They are usually defined by some kind of mathematical formula so their results are more stable and can be obtained again by performing the same measurement. However, it is hard to design an objective method that would measure the overall perception of human ear.

1.2 Overview of the Thesis

This thesis is written within an artificial bandwidth expansion project that was a collaboration with the Laboratory of Acoustics and Audio Signal Processing of Helsinki University of Technology and the Speech and Audio Systems Laboratory of Nokia Research Center. The project involved developing new features to an already existing expansion algorithm that was developed in a previous collaboration project and running subjective listening tests. The emphasis in the algorithm development was on the processing of fricatives and on the classification of sounds. The target was to develop an algorithm that would be robust and suitable for applications of mobile communication systems.

During the last decade, different algorithms for artificial frequency expansion have been developed. The main goal for all of them has been the same, to improve the speech quality by adding new spectral components to bandlimited speech. There is a variety of different ways to approach this problem and these methods are discussed in **chapter 2**.

Standardized quality assessment methods, both subjective and objective, are reviewed in **chapter 3**. Also some aspects that has to be taken into account when selecting a suitable assessment method are discussed.

The artificial bandwidth expansion algorithm that was further developed within this thesis, is presented in **chapter 4**. For analyzing the quality of the implemented ABE algorithm, a listening test was arranged. The description of test arrangements and the results of the test are presented in **chapter 5**. **Chapter 6** completes the thesis with analysis of listening test results and conclusions of the topic.

Chapter 2

Methods of Artificial Bandwidth Expansion

Artificial bandwidth expansion is a relatively new concept; publications covering the topic are mainly from the last decade. Different approaches to the problem exist and promising results have been obtained from many of them. However, so far none of the methods has proven superior to others which is why the topic is highly interesting and challenging.

In this chapter, artificial bandwidth expansion methods for narrowband telephone speech of bandwidth 300–3400 Hz are discussed. They are divided into two groups. Methods of the first group are based on envelope aliasing or nonlinear processing. Signal processing techniques behind them are quite simple and they only try to add some energy to the higher band in order to enhance the speech quality. Methods of the second group are based on a linear model of human speech production. The starting point for these algorithms is different from algorithms of the first group since they try to simulate accurately high frequencies of speech and reconstruct the wideband signal.

Artificial bandwidth expansion methods try to expand the telephone bandwidth, 300–3400 Hz to cover frequencies up to 7 kHz or 8 kHz. Some of the algorithms also add spectral components to cover the frequency band below 300 Hz. In this chapter, artificial expansion methods towards high frequencies are discussed.

2.1 Envelope Aliasing and Nonlinear Processing

Yasukawa has proposed a frequency expansion method that is based on envelope aliasing [Yas94]. The block diagram of the algorithm is in figure 2.1. The narrowband signal (sam-

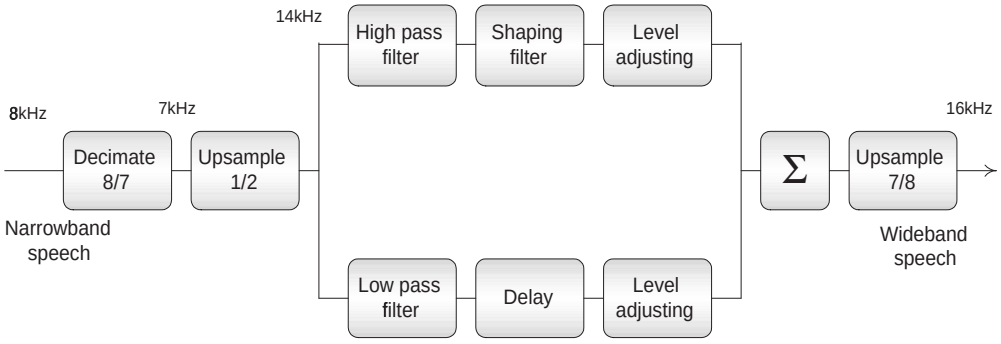

Figure 2.1: Block diagram of the frequency expansion algorithm based on envelope aliasing designed by Yasukawa [Yas94]

pling frequency $f_s = 8kHz$) is upsampled and as a result the aliased frequencies appear in the highband. The sampling frequency is now 14 kHz and a mirror image of the narrowband signal appears in the frequency range of 3.5-7kHz. The signal is highpass filtered with cut-off frequency $f_c = 3.5kHz$. The spectral content is then shaped by a shaping filter and the energy of the signal is adjusted to a suitable level compared to the original narrowband signal. The shaping filter is a linear phase FIR lowpass filter with cutoff frequency $f_c = 2kHz$ and stopband edge $f_{sb} = 8kHz$. Some level adjusting is needed for setting the energy level of the highband to a suitable level. In order to construct the wideband signal, the shaped highband and the original narrowband signal are summed and finally the summed signal is converted to 16kHz with 7-to 8- conversion.

If the original signal was in the first place upsampled to 16 kHz, there would have been a 1200 Hz gap in the spectrum around 4 kHz after aliasing because the actual frequency range of telephone speech is limited to 3.4 kHz. By using a sampling frequency of 14 kHz in aliasing, the gap is smaller (200 Hz) and the signal can be upsampled to 16 kHz afterwards. As a result the frequency region of the expanded signal is from 300 Hz to 7 kHz.

Yasukawa has also proposed a method that utilizes nonlinear processing [Yas96]. The block diagram of this is shown in figure 2.2. The main idea behind this algorithm is almost the same as in envelope aliasing, but now new spectral components are produced using rectification. As a result, harmonics of the original signal appear in the highband. The spectral content of the highband is modified using a shaping filter that has the same characteristics as in the previous algorithm. The level of the highband is adjusted to a suitable level and the overall result is constructed by summing the highband signal and the original narrowband signal together.

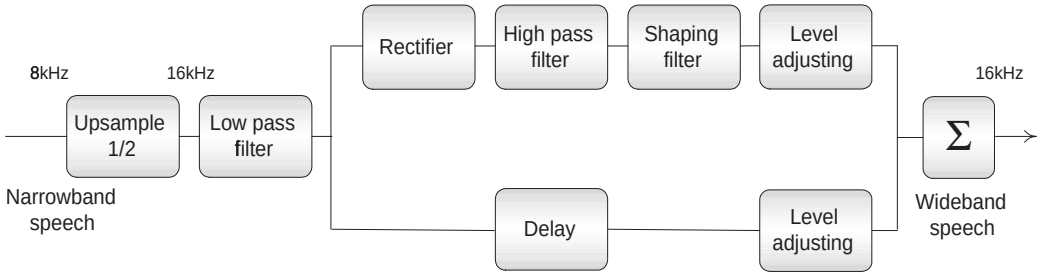

Figure 2.2: Block diagram of the frequency expansion algorithm based on rectification designed by Yasukawa [Yas96]

Both of Yasukawa's expansion algorithms are based on filterings and resamplings. The performance of the first method which is based on aliasing, is strictly dependent on the characteristics of the shaping filter. Spectral envelopes of different speech sounds differ from each other in many ways and that is why one single shaping filter can not be optimal for every sound. Added frequency components are a copy from narrowband and therefore they are not especially designed for the highband. In the algorithm that uses rectification, new spectral components are harmonics of those in the narrowband and therefore they are justified. But also in this algorithm, the shaping filter plays an important role in the performance.

2.2 Methods Based on Linear Model of Human Speech Production

Many of the bandwidth expansion algorithms are based on a source-filter model of human speech production which is presented in figure 2.3. In the model, an excitation signal is filtered with a filter that contains the vocal tract information. There are two excitation signals, one for voiced sounds and another for unvoiced sounds. The former corresponds to glottal pulses and the latter to the turbulent air-flow that originates from a constriction of the vocal tract. The vocal tract filter alters the frequency content of excitation signals and by changing its shape during speech production different phonemes can be produced.

The creation of the highband of speech using the source-filter model is illustrated in figure 2.4. Excitation signal describes the fine structure of the spectrum and is usually obtained using the existing information of the narrowband signal. It is a residual signal that results from filtering the narrowband signal with an analysis filter which is the inverse filter whose impulse response is given by auto-regressive (AR) coefficients $a(k)$. The result, the residual signal, is upsampled by two and aliased frequencies are lowpass-filtered. Finally the

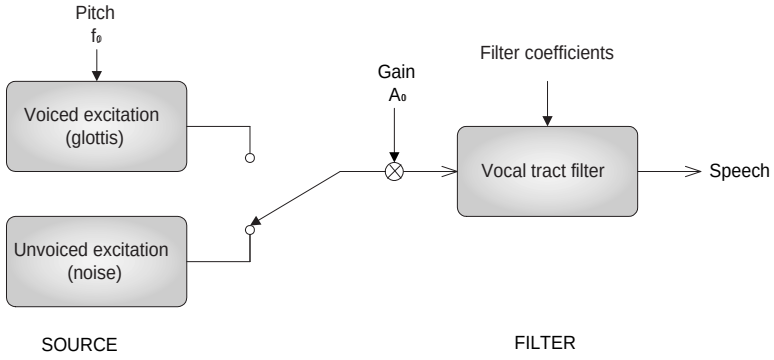

Figure 2.3: A source-filter model of human speech production system

upsampled residual is filtered with synthesis filter that predicts the wideband envelope. The overall result, artificially expanded signal is the output of synthesis filtering or a combination of the original narrowband signal and the highband of the output of the synthesis filtering. The creation of the vocal tract filter which models the spectral envelope is the main problem in implementing this kind of artificial expansion algorithm.

Different approaches for bandwidth expansion methods based on the linear model of human speech production exist and they are discussed next.

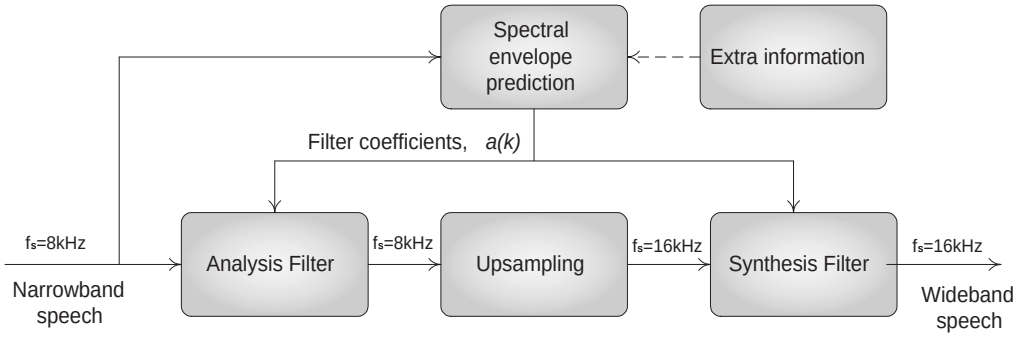

Figure 2.4: Block diagram of how to produce expanded speech using the source-filter model

2.2.1 Codebook Mapping

In codebook mapping, the spectral envelope is selected from a pre-designed codebook [CH96] [CH97] [EK99] [EH99] [Yas98] [JV00]. The narrowband envelope is compared to possible wideband envelopes of the codebook and the one closest to the narrowband envelope is selected.

Spectra are usually stored as mel-frequency cepstral coefficients, AR parameters or as LSPs (Line Spectral Pairs) and they are often encoded using vector quantization (VQ). The creation of codebooks is done with some training algorithm and training data. In the literature, an algorithm called LBG training algorithm (named by its inventors, Linde, Buzo and Gray) is often used. Training data is selected so that it contains wideband speech frames that are as diversified as possible. For example the TIMIT speech database is often used.

Different criteria for clustering and for selecting the closest envelope from a codebook are presented in the literature.

Enbom [EK99] uses a codebook which contains narrowband mel frequency cepstral coefficients (MFCC) and the corresponding wideband spectral envelope stored as LSP-coefficients. The codebook structure can be seen in figure 2.5. The criterion for clustering and for selecting the best codeword from the codebook is the following:

$$D_n = \sum_{i=1}^{K} (c(i) - \hat{c}_n(i))^2, \tag{2.1}$$

where $c(i)$ is the MFCC of the current speech frame, $\hat{c}_n(i)$ is the MFCC of codeword n and K is the order of MFCC. The selected codeword is the one that achieves the smallest measure.

1	narrowband MFCC	wideband LSP
2	...	...
3	...	...
4	...	...
⋮	⋮	⋮
N	...	...

Figure 2.5: Possible codebook structure [EK99]

The single codebook of Chan and Hui [CH97] consists of wideband spectrum LSP vectors. They choose the best codeword on the grounds of the spectral distortion measure which is calculated only from the lowband and defined as:

$$SD_{AV} = \left[\frac{1}{N} \sum_{n=1}^{N} \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} (10 \log |H_n(\omega)| - 10 \log |\hat{H}_n(\omega)|)^2 d\omega \right]^{\frac{1}{2}}. \tag{2.2}$$

$H(\omega)$ is the LPC spectrum of the current speech frame and $\hat{H}(\omega)$ is the LPC spectrum of the codeword. Again the idea is to minimize the measure.

Jax and Vary generate also a single wideband pre-trained codebook that contains the wideband AR coefficients for estimating the wideband spectral envelope [JV00]. The basis for the codebook search method is a hidden Markov model (HMM) of the speech generation process. Certain features of the narrowband signal are first extracted. The feature vector consists of eight cepstral coefficients, the normalized frame energy and the gradient index. The extracted features are compared with a statistical model of the speech production process using hidden Markov model. A suitable estimate $\hat{\mathbf{C}}$ for the AR coefficients $\mathbf{C}$ is calculated by minimizing the squared error between the estimate and true coefficients.

The benefit of the algorithms developed by Jax and Vary [JV00] or Chan and Hui [CH97] is that only one codebook is needed. Therefore, the AR coefficients used for the analysis filter $H_A(z)$ and for the synthesis filter $H_S(z)$ are the same and their transfer functions are mutually inverse:

$$H_S(z) = \frac{1}{H_A(z)}. \tag{2.3}$$

The basic codebook mapping can be varied in many ways. Epps and Holmes [EH99] present a method that interpolates between several envelopes. In codebook mapping with interpolation, N closest narrowband envelopes are selected from the codebook and the corresponding highband envelopes are interpolated. A codebook can also be split into several parts. For example separate codebooks can be used for voiced and unvoiced sounds. This is justified because different degrees of voicing are associated with different spectral envelope shapes [EH99]. While training separate codebooks, the training data has to be divided into corresponding parts and the expansion algorithm has to be able to detect and classify sounds.

When the envelope is selected from a codebook, the number of possible envelopes is limited to the size of the codebook. The codebook should cover all the possible phonemes and also different ways to pronounce them. Therefore an extensive codebook would require a large amount of computational capacity. At the same time the distortion is dependent on how well the selected envelope predicts the real highband. Even a single incorrectly selected envelope might cause a severe degradation of speech quality which is why a variety of refinement methods have been used to minimize occasional mismatches. Enbon et al. [EK99] propose two methods for this purpose. The first one is called envelope smoothing and through it rapid changes in the spectrum are avoided by forming the wideband envelope as a weighted sum of last three chosen codewords. The second method is used to avoid overestimates

of the highband signal power by marking some of the frames “dangerous”. If any of the “dangerous” envelopes is selected during a voiced sound, its power is lowered with 10 dB.

2.2.2 Statistical Methods

Statistical methods to generate the wideband spectral envelope are based on statistical dependences between the narrowband and highband spectra.

Cheng, O’Shaughnessy and Mermelstein [COM94] have developed a statistical recovery function to recover wideband speech from the narrowband speech. They segment speech signals into random sources that have probability density functions and are characterized by mean and covariance matrices. Each narrowband speech frame is generated by a combination of N random sources and each highband speech by a combination of M random sources. In the creation of the highband (figure 2.6), weight factors for all the random sources are evaluated from a pre-trained codebook. The highband is then constructed using a filter bank with weight factors where each filter represents a random source’s autoregressive spectrum in the highband. The outputs of the filter bank are summed and the highband of the constructed signal is combined with the original narrowband signal.

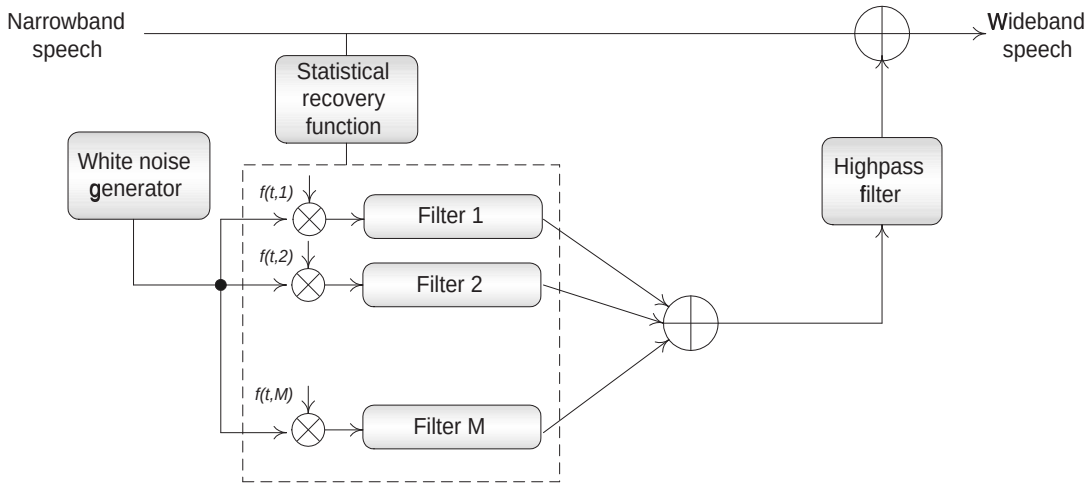

Figure 2.6: Block diagram of a wideband recovery system based on statistical recovery function [COM94]

Hosoki, Nagai and Kurematsu present a frequency expansion algorithm based on subband hidden Markov model [HNK02]. Their idea is to divide the wideband speech signal into a number of subbands and extract their features independently. First the wideband speech frames are modeled with feature vectors which are combined and the HMMs are trained by

the EM algorithm. In the learning phase the HMM learns to model the feature of the signal as a single subband and the relationship of the corresponding subbands. After that each HMM is divided into subband HMMs and an energy HMM. In the reconstruction phase an observed narrow band signal is decoded by the Viterbi algorithm.

Park and Kim [PK00] have proposed a method where the wideband envelope is obtained using Gaussian mixture model (GMM). The block diagram is shown in figure 2.7. The parameters of the GMM are estimated using a set of training speech. Spectral envelope, pitch and power information are first extracted from a narrowband signal. A conversion of the envelope from narrowband to wideband is made using GMM. The wideband speech is synthesized using LPC and the power adjusting is carried out using a pre-trained codebook that contains power ratios of the narrowband and highband speech. Finally the highband of the synthesized signal is combined with the original narrowband signal. Park and Kim have evaluated the performance of their method using both objective and subjective methods. The objective measure was Euclidean distance of LPC cepstrum and compared to conventional codebook mapping, the GMM-method obtained slightly better results.

Cheng [COM94] et al. have reported that most of the highband speech was successfully reconstructed but there were problems with fricatives like /s/ and /f/. They are problematic because their information is mainly in the highband and there is a possibility that the codebook search algorithm might make a wrong classification. Also Park and Kim [PK00] conclude that the reconstructed speech is somewhat noisy but still the objective and subjective tests gave relatively good results.

2.2.3 Linear Mapping

Linear mapping has been used in spectral envelope prediction although it is considered as a non-linear problem [AHW95] [EH99] [CGMS01]. Epps and Holmes [EH99] present the basic idea of linear mapping. The input narrowband envelope is characterized by a vector of parameters $\mathbf{x} = [x_1 x_2 \dots x_n]$ and the wideband envelope by another vector $\mathbf{y} = [y_1 y_2 \dots y_m]$. The linear mapping is then simply denoted as

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{2.4}$$

where $\mathbf{W}$ is determined using least squares. For this purpose, matrices $\mathbf{X}$ and $\mathbf{Y}$ are constructed from a database so that they contain narrowband and highband vectors. Then $\mathbf{W}$ can be calculated from:

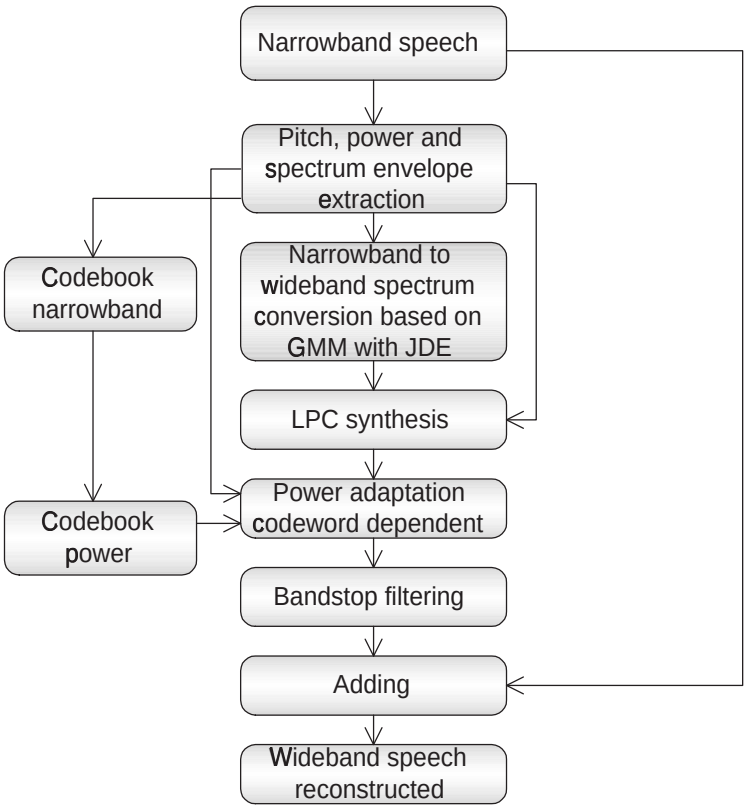

Figure 2.7: Block diagram of the frequency expansion algorithm developed by Park and Kim [PK00]

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{2.5}$$

Avendano, Hermansky and Wan take advantage of linear mapping by filtering time trajectories of LPC-cepstral coefficients of narrowband signal through a multidimensional filter designed on training data [AHW95]. A bank of multi-input single-output filters is designed using time trajectories of cepstral coefficients which are calculated from autoregressive coefficients. Each filter of the bank is designed to map a time window of the eight coefficient trajectories from narrowband frame to a particular coefficient corresponding to the current wideband frame. Therefore the output of each filter is:

$$\hat{C}_r^d(k) = \sum_{i=1}^{p} \sum_{l=-M}^{M} W_{i,r}(l) C_i(k-l), \tag{2.6}$$

where $C_i(k)$ $(i = 1, 2, \ldots, p)$ is the ith cepstral coefficient of AR coefficients $a(k)$, $\hat{C}_r^d$ is

the estimate of the rth cepstral coefficient corresponding to the envelope of wideband signal and $W_{i,r}$ are FIR filter coefficients (found such that $\hat{C}_r^d$ is the least squares estimate of the original C_r^d).

Avendano et al. report some over-estimates of the spectral envelope slope at high frequencies for voiced frames. This causes perceivable artifacts which can be very annoying. The problem is that it is hard to find a linear solution for a non-linear problem. It is probable that a linear solution in this case results in either over-estimates for voiced sounds or in under-estimates for unvoiced sounds because of different spectral characteristics of vowels and fricatives.

2.3 Possibilities and Limitations

Although the same physical acoustic configuration generates both the narrowband and the highband of speech signal, an interesting question is that is the information that can be extracted from narrowband speech enough for creating the highband?

The information theoretic background for artificial bandwidth expansion algorithms has been examined in a few publications. In the first paper, written by Nilsson, Andersen and Kleijn [NAK00], a method for estimating a lower bound on the mutual information (the reduction in uncertainty of a random variable due to another random variable) between a low band spectral coefficient vector and a high band spectral slope and gain was proposed. Their results show that there is mutual information between the low and high frequency band. For the slope, the mutual information is approximately 0.1 bit and for the gain it is approximately 0.45 bit.

In the second paper, Nilsson, Gustafsson, Andersen and Kleijn have studied the dependency between the spectral envelopes of speech in disjoint frequency bands of 0.3–3.4 kHz and 3.7–8.0 kHz [NGAK02]. The envelopes were modeled with a Gaussian mixture model (GMM) based on mel-frequency cepstral coefficients and the log-energy-ratios of the frequency bands. Mel-frequency cepstral coefficients modeled the shape of the envelopes and log-energy-ratios modeled the energies. In the study two dependency measures were used, mutual information and differential entropy.

The results of the experiments made by Nilsson et al. show relatively low dependency between the two bands and the authors claim that an expansion method that uses a mapping between spectral shapes in the narrow- and highband will most likely result in a perceivable spectral error in the highband. As a conclusion they suggest that a reasonable expansion method uses memory-less mapping and tries to expand the signal such that it sounds pleas-

ant instead of trying to predict the true highband.

According to Jax and Vary [JV02] a frequency expansion algorithm that operates without any side information of the missing highband has to exploit mutual dependencies between the available and missing frequency bands of the speech signal. In their paper, they have studied the relationship between the maximum achievable quality of a bandwidth expansion algorithm and the mutual information between the bandlimited speech and the missing highband from an information theoretic perspective. For evaluating the performance of an frequency expansion algorithm, Jax and Vary measure the spectral distortion of the spectral envelope of the expanded signal with respect to the original wideband signal. Since the narrowband information of the estimate and the original signal are the same, only the missing frequency range is measured. The measure is a log spectral distortion (LSD) and defined as:

$$d_{LSD}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(20\log_{10}\frac{\sigma_{rel}}{|A_{mb}(e^{j\omega})|} - 20\log_{10}\frac{\tilde{\sigma}_{rel}}{|\tilde{A}_{mb}(e^{j\omega})|}\Big)^2 d\omega, \tag{2.7}$$

where $A_{mb}(e^{j\omega})$ and σ_{rel} denote the modeled frequency spectrum and relative gain of the missing frequency band of the original wideband signal, respectively, and $\tilde{A}_{mb}(e^{j\omega})$ and $\tilde{\sigma}_{rel}$ are the corresponding parameters of the artificially expanded signal. The LSD can also be expressed in the cepstral domain and after taking the root mean-square (RMS) of the LSD, the formula for the quality measure is:

$$\bar{d}_{LSD} = \frac{\sqrt{2}10}{\log_e 10}\sqrt{E\Big\{\frac{1}{2}(c_0-\tilde{c}_0)^2 + \sum_{i=1}^{\infty}(c_i-\tilde{c}_i)^2\Big\}}, \tag{2.8}$$

where the cepstral coefficients c_0, c_1, ...are calculated from the AR coefficients and the relative gain.

Jax and Vary present a lower bound for the above mentioned distortion measure $\bar{d}_{LSD}$ with respect to the mutual information and the entropy:

$$\bar{d}_{LSD} \geq C_d e^{\frac{1}{d}(h(\mathbf{y}-I(\mathbf{x};\mathbf{y})))} \tag{2.9}$$

where $\mathbf{x}$ contains the features of narrowband signal, $\mathbf{y}$ is the vector of features of the missing frequency band, $C_d = \frac{\sqrt{2}10}{\log 10}\sqrt{\frac{d}{2\pi e}}$ and d is the dimension of $\mathbf{y}$. The lower bound can be applied to any linear or non-linear memory-less frequency expansion algorithm. It is a theoretic bound and the result does not point out how the optimal expansion method should be implemented. As the lower bound for the LSD depends on the mutual information

$I(\mathbf{x};\mathbf{y})$, the selection of the elements in $\mathbf{x}$ affects the result. Therefore, it would be desirable to select a vector that maximizes the mutual information.

As a conclusion of these three papers [NAK00], [NGAK02],[JV02], it can be said that there is a low limit for how much we can know about the highband envelope on the basis of the narrowband spectral envelope only. However, because the lower bound of [JV02] can be applied only to memory-less algorithms it does not take into account that usually there is more information available than just one single speech frame. Extra information can be obtained, for example, by examining the whole context of a frame. In addition, the objective quality measure like LSD does not completely correspond to the perceived speech quality. The speech quality can be enhanced by adding some frequency components to the highband without even trying to mimic original spectral content of higher frequency band.

Chapter 3

Evaluation Methods for Speech Quality

3.1 Introduction

Sound quality is an abstract concept which has no specific definition. It can mean different things in different situations. People evaluate quality using a scale good vs. bad, by describing quality with adjectives or by comparing it to some known situation. Often the idea is actually to measure quality differences and in the case of speech quality, the reference signal is usually natural speech. On the other hand in some applications, the only quality requirement for transmitted speech is that it is intelligible, i.e. it is understood what is been said.

There are several issues underlying speech quality. The first one is speech intelligibility. It is dependent on the speaker, the transmission channel and the listener. In other words the quality of speech in a telecommunications system is a sum of many factors that are dependent not only on the system itself but also people using it. For example, naturalness and sound quality are usual characteristics of a transmission system whereas listening effort, conversational effort and expectations are related to the user. If the speech has a high intelligibility, like in many high quality communication systems, the next step is to evaluate the identifiability of the speaker and the naturalness of the speech.

Subjective listening tests, which are tests carried out with people, take into account both human and technical influences. No matter how carefully subjective tests are implemented the results are always slightly different, because every listener pays attention to different things. Objective assessment methods, on the other hand, are quality assessment techniques

based on modeling of human auditory system and they lead to more stable results. However, usually they measure only one feature of the sound, not the overall quality.

Subjective methods for evaluating speech quality and the selection of a proper method are discussed in section 3.2. Section 3.3 reviews some common objective measures.

3.2 Subjective Methods

Due to the rapid development of telecommunications systems the standardization of digital speech coding and processing has become increasingly important. This has led to a need to understand and develop subjective quality assessment methods in order to evaluate systems before they are released onto the market [Dim91]. From the commercial perspective the idea is to determine whether or not a product meets design criteria and customer expectations. Standardization of assessment methods helps to minimize the variability caused by a physical environment and circumstances of a test situation. Results of listening tests become more consistent and valuable and the comparison and grading of systems is easier and more reliable [Too82].

3.2.1 Factors Affecting the Selection of a Test Method

The degradation of the speech quality can depend on many impairment factors. Therefore, classification of them can help in finding answers to the questions like why speech sounds like it does and which quality assessment method should be selected.

According to Richards [Ric73], impairment factors that cause quality degradation can be divided into three groups. The first group consists of factors that cause increase in difficulty to listen. The second group includes factors that cause difficulty in talking and finally the last group is formed by impairment factors that complicate conversation. All of these impairment factors can occur alone or together depending on the communication system. For example communication channels can cause various impairment factors. Quantization noise is an example of impairment factors of the first group, it complicates listening. Long telephone lines can cause impairment factors that belong to the second group. If the speakers own voice is reflected back to the speaker, it is hard to talk on the telephone. Another example of impairment factors is noise in a room where the test is taking place. It is an example of an influence of the environment. It can cause difficulties in hearing so it belongs to the first group of impairment factors. Also impairment factors caused by unfamiliar accents or hearing difficulties belong to the same group

The most important criterion in the selection of a suitable assessment method is how the results are to be analyzed and used. Also the environment of the test situation and participants themselves must be taken into account in designing listening tests [Dim91]. As Toole ([Too82]) has stated:

> "A number of physical factors can distort the sound at the listeners' ears, and psychological factors can distort their perceptions. Controlled listening tests reduce the influence of these factors."

The audio signal that is under evaluation should be controlled carefully. The sound source can be the one that produces the sound, a flute, a car, a human been etc. The sound can also be produced through loudspeakers or headphones as is often the case when evaluating the quality of a communication system. Different spatial impressions can be created using loudspeakers. On the other hand the spatial impression can be minimized using headphones. In addition, it is possible to give different responses to both ears through headphones. The sound field where the listener is situated affects the result as well. Different standardized places, like anechoic chamber, reverberation chamber and listening room exist and they are often used in subjective listening tests.

Listeners of a listening test should be selected carefully. If the idea is to get an extensive picture of how people in generally react with some system, the listeners should be selected randomly from possible end-users. The use of research engineers as listeners can be justifiable in some cases since they are the ones that plan and develop systems. And if the target is to develop a high-quality system, the listeners should be experienced ones who can detect even the smallest faults.

Other factors that have an influence on the selection of a listening test method are cost and expected quality of the system under evaluation [Dim91]. Firstly, a listening test that involves numerous subjects or requires extensive arrangements, can be very expensive. Depending on what the resources are and what is the purpose of the test, the cost has to be taken into account. Secondly it is possible to detect impairment factors of different size with different assessment methods. For example using some grading scale for evaluating a high quality system may concentrate on the best grade, whereas the paired comparison test would disclose small differences much better.

3.2.2 Listening Tests

According to Dimolitsas [Dim91] subjective listening tests can be divided into four groups:

- Listener opinion tests
- Articulation and diagnostic techniques
- Conversation opinion tests
- Field tests

Listening opinion tests are subjective measures of naturalness. International Telecommunication Union (ITU) has made recommendations [P.896] that cover these tests. Articulation and diagnostic tests are measures for speech intelligibility. Particularly with low bit rate transmission it is enough that speech is intelligible, so the target is to find out how well this goal is achieved. The third group, conversational opinion tests, includes tests that are employed when analyzing how the system behaves in conversation situations. These tests are usually much more complicated to implement. Field tests are usually conversation tests that are as realistic as possible. The target is to examine how all the environmental conditions and the overall system affect the system under evaluation.

Listener Opinion Tests

Opinion ratings and paired comparison tests are examples of listener opinion tests.

Opinion rating tests measure the degree of speech-quality satisfaction or how well some feature differences can be detected. Various five-point category-judgement scales are often used for evaluating sound quality. The scale should be selected carefully and the layout and wording as seen by subjects in experiments is very important.

In an absolute category rating (ACR) test a listener hears a sample reproduced over the system under evaluation. The listener is asked to express an opinion in terms of a chosen scale. According to ITU-T the most frequently used scale in ACR tests is so called MOS (mean opinion score) scale [P.896]. A MOS scale for listening-quality is presented in table 3.1.

For evaluating listening-effort and loudness preference with MOS scale, there are specific verbal descriptions for each score as well. The scales are denoted as MOS_{LE} and MOS_{LP} and they are presented in tables 3.2 and 3.3.

From these five-point grades weighted mean value is calculated to obtain Mean Opinion Score (MOS). The MOS method has some certain advantages. Firstly, it is quite easy to implement, because trained listeners are not usually required. It is also possible to assess different impairment factors simultaneously because the overall naturalness is examined all the time.

Score	Quality of the speech
5	excellent
4	good
3	fair
2	poor
1	bad

Table 3.1: MOS scale for listening-quality

Score	Effort required to understand the meaning of sentences
5	Complete relaxation possible; no effort required
4	Attention necessary; no appreciable effort required
3	Moderate effort required
2	Considerable effort required
1	No meaning understood with any feasible effort

Table 3.2: MOS scale for listening effort

A good way to obtain information on the detectability or some analogous property of a sound as a function of some objective quantity is to use a quantal-response method [P.896]. The subject is asked to vote on a scale that is presented in table 3.4. The alternative "B" is understood to mean "Detectable but not objectionable". Opinion scores are usually not well suited for this kind of tests because objectionability and intelligibility differ from detectability both in kind and in degree.

Degradation category rating (DCR) affords higher sensitivity than ACR tests. The test signal is compared with a reference signal and the degradation in quality is graded using a five-grade scale known as DMOS (degradation MOS) scale, which can be seen in table 3.5. The DCR procedure is suitable for evaluating good quality speech.

One of the disadvantages of opinion ratings is that test conditions affect the result. Also small differences in signals do not necessarily show in results, because a listener might give same grades for signals that are very close to each other.

In a paired comparison test a signal pair is presented to the listener. The listener is asked to decide which one is better, signal A or signal B. From the results a preference score is defined as the percentage of how often the listener chooses the concerned signal as the preferred signal.

Score	Loudness preference
5	Much louder than preferred
4	Louder than preferred
3	Preferred
2	Quieter than preferred
1	Much quieter than preferred

Table 3.3: MOS scale for loudness preference

Score	Detectability
A	Objectionable
B	Detectable
C	Not detectable

Table 3.4: Detectability opinion scale.

It is also possible to use a scale when comparing the quality of two sound samples with each other. The listener is asked to give his or her opinion on the quality of the second to the quality of the first sound sample using a scale that is shown in table 3.6. In ITU-T recommendations this test is called comparison category rating (CCR) [P.896].

Paired comparison is a good choice for a test, if it is needed to evaluate delicate and precise differences. It is more sensitive to them because the assessment is relative. But if impairment factors of the signal pair are considerably different, it might be difficult to judge which one is better. In addition, the paired comparison test without any grading does not tell how big the difference is between the two signals, the only thing it tells is that which one is better.

There are many different variations of these basic tests. All of them have their own advantages and disadvantages. For example, in a test called order ranked preference designs a test signal is compared against some other signal (but not a reference signal). The idea is to rank the signals from the best to the worst. The same test can also be implemented with various test signals. In these higher order ranked preference designs the listener has to rank a set of signals (usually three or four).

Score	Degree of degradation
5	degradation is inaudible
4	degradation is audible but not annoying
3	degradation is slightly annoying
3	degradation is annoying
1	degradation is very annoying

Table 3.5: DMOS scale for analyzing the degree of degradation.

Score	Quality of second compared to the quality of first
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

Table 3.6: Scale for comparing qualities of two sound samples.

Articulation and Diagnostic Techniques

According to Karjalainen [Kar00] speech intelligibility is dependent on three factors: how well the speaker is able to produce speech of good quality and understandable message, how well the channel is able to transmit the message and how well the listener is able to receive and analyze it. Looking from a technical aspect the second factor is usually the most interesting one and often the concept “speech intelligibility” of a system refers to it. The next two assessment methods are subjective ways to measure it.

Articulation tests measure the percentage of correctly recognized speech sounds in the receiving end of a system. Sounds can be monosyllables (syllabic articulation) or phonemes (sound articulation). It is also possible to measure word intelligibility, which is a ratio of sent words to correctly received words [KN88]. This kind of word intelligibility assessment method takes into account the whole context of concatenated vocal sounds instead of individual phonemes.

Diagnostic rhyme test (DRT) is a two-choice test where a usually trained listener hears a word that is one of the two given words that have different initial consonants with respect

to their phonetic characters. The listener has to decide which one the heard word was. Different word pairs assess different aspects of phonemic articulation.

Conversation Opinion and Field Tests

Listener opinion tests do not always give enough information on the system under evaluation because test situations are artificial. Conversational opinion tests are one step closer to the real word, where most of the voice communication is conversation between two or more people. The test simulates a real conversational situation, so the speech quality of the overall system is under evaluation. Conversational opinion tests can be organized in a laboratory environment or on a field.

Laboratory conversation opinion tests are suitable when some impairment factor in the link is limiting the quality of speech [Dim91]. It is possible to assess effects of different conditions and to find out more information on the impairment factor. They are held in controlled environment in order to be able to specify where the problem is.

Field tests are suitable when the behaviour of the whole end-to-end connection is not known exactly or when it is not possible to create a system that corresponds to it perfectly [Dim91]. Field tests are supposed to be as realistic as possible.

3.3 Objective Methods

Objective quality assessment methods have a universal meaning if they meet standards and are implemented with calibrated equipment. When compared to subjective quality assessment methods, objective methods are usually cheaper and easier to implement, which is a considerable advantage. However, it is hard to find an objective measure that would completely replace subjective methods although an objective method can at its best correspond with subjective methods.

Signal-to-noise ratio (SNR) is probably the most often used objective measure. The basic SNR is defined as:

$$SNR = 10 \log_{10} \frac{x^2(x)}{[x(n) - x_d(n)]^2}, \tag{3.1}$$

where $x(n)$ is a clean signal and $x_d(n)$ is the signal after the system under evaluation. The usability of SNR as a speech quality measure is very limited. It is not a good measure of perceptual quality because it is very sensible to all the differences between the signals

although they might not be essential for perceived speech quality.

A closely related measure for plain signal-to-noise ratio is segmental signal-to-noise ratio. The signal is divided into small frames, usually 15-30 ms and SNRs are calculated separatetly for all of them. Segmented SNR is then obtained by averaging the ratios:

$$SNR_{seg} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=m_j-N+1}^{m_j}\frac{x^2(n)}{\hat{n}^2(n)}\right], \tag{3.2}$$

where M is the number of frames, $x(n)$ is the clean signal and $\hat{n}(n)$ is noise.

Spectral distance measures are often used in comparing two signals. The fact that spectral differences represent the signal much better than time domain properties from the hearing point of view, speaks for spectral distance measures. For example Epps and Holmes [EH99] use the following spectral distortion measure to measure spectral distortion between two envelope shapes:

$$D_{HC} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{0.25\omega_s}\int_{0.25\omega_s}^{0.5\omega_s}\left[20\log_{10}\left(G_C\frac{A_k(\omega)}{\hat{A}_k(\omega)}\right)\right]^2 d\omega} \tag{3.3}$$

where

$$G_C = \frac{1}{0.25\omega_s}\int_{0.25\omega_s}^{0.5\omega_s} 20\log_{10}\left(\frac{A_k(\omega)}{\hat{A}_k(\omega)}\right)d\omega, \tag{3.4}$$

$A_k(\omega)$ and $\hat{A}_k(\omega)$ are the original envelope and the envelope of the signal after the system under evaluation of kth frames and ω is the sampling frequency.

Several distortion measures that take into account the knowledge of the human auditory system have become common in recent years. They measure the perceived speech quality by trying to give more weight to distortions that are most significant for human ear and less weight to those that are inaudible or nearly so. Just to mention one, Perceptual Evaluation of Speech Quality (PESQ) is one of this kind of measures.

Chapter 4

Algorithm

4.1 Introduction

An artificial bandwidth expansion (ABE) algorithm has been developed in the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology. The first version was developed by Jutta Mahkonen in her master's thesis in 1999 [Mah99]. The second version was made by Kimmo Käyhkö in 2001 [Käy01], which was also the starting point for this version. The ABE algorithm has been simulated using MATLAB programming environment.

Originally the idea behind the ABE algorithm was to generate an algorithm that would be applicable to telephone systems. It was to be an algorithm that would improve, and in no circumstances degrade speech quality. The idea was to increase the sampling frequency from normal $8kHz$ to $16kHz$, to add new frequency components to the higher band of the signal and to leave the original narrowband untouched. That way the bandwidth of the speech would increase from 0.3–3.4 kHz to 0.3–7.7 kHz. In addition, the ABE algorithm was to be developed so that it would be possible to implement it in real-time.

In a telecommunication network a potential platform for ABE would be a communication terminal or it could be implemented on the network side. For both applications, the algorithm should be written for some DSP processor.

The algorithm is based on aliasing which is carried out by inserting a zero between every sample. In the frequency domain this means that the bandwidth is expanded by creating a mirror image of the narrowband spectrum to the highband. Aliasing generates new spectral components which distorts the naturalness of sound. Therefore the spectrum has to be modified further in order to achieve the main goal i.e. enhance the speech quality.

Within this thesis, the goal of the algorithm development was to develop it further by improving the classification and processing of different speech sounds, particularly fricatives and stop-consonants. Moreover, the algorithm for recognizing fricatives was to be simple and robust.

4.2 Description of the ABE Algorithm

The first version of the ABE algorithm was developed by Jutta Mahkonen in her master's thesis in 1999 [Mah99]. The second version was made by Kimmo Käyhkö in 2001 [Käy01] which is also the starting point for this third version.

4.2.1 ABE Algorithm in Short

A block diagram of the ABE algorithm is shown in figure 4.1. The narrowband speech signal is divided into frames. The default value for the size of the frame is 20 ms. Each frame is windowed with a Hamming window of length 30 ms so it overlaps 5 ms with both of its adjacent frames.

The frequency expansion is carried out by up-sampling the narrowband signal frame by two. Then the FFT is calculated and the spectrum is modified with a specific attenuation or amplification function.

As a result of the aliasing, the harmonic structure of the low band spectrum appears in the high band as well. This causes unpleasant effects for voiced sounds. Therefore spectral peaks of the high band are smoothened by averaging.

The expanded and averaged frame is transformed back to the time domain by using inverse fast fourier transform, IFFT. The effect of the Hamming window at the beginning of the algorithm is compensated by unwindowing the frame with inverse Hamming window. After that all the frames are cascaded. In order to cascade the frames nicely and to avoid generation of clicks in the signal, the boundaries of the frames are smoothed by time-domain averaging.

4.2.2 Framing and Aliasing

The ABE algorithm starts to process the narrowband ($f_s = 8kHz$) input signal by partitioning it into frames. By processing the signal in short frames, it is possible to operate in the frequency domain which makes possible the control of frequency components. Moreover,

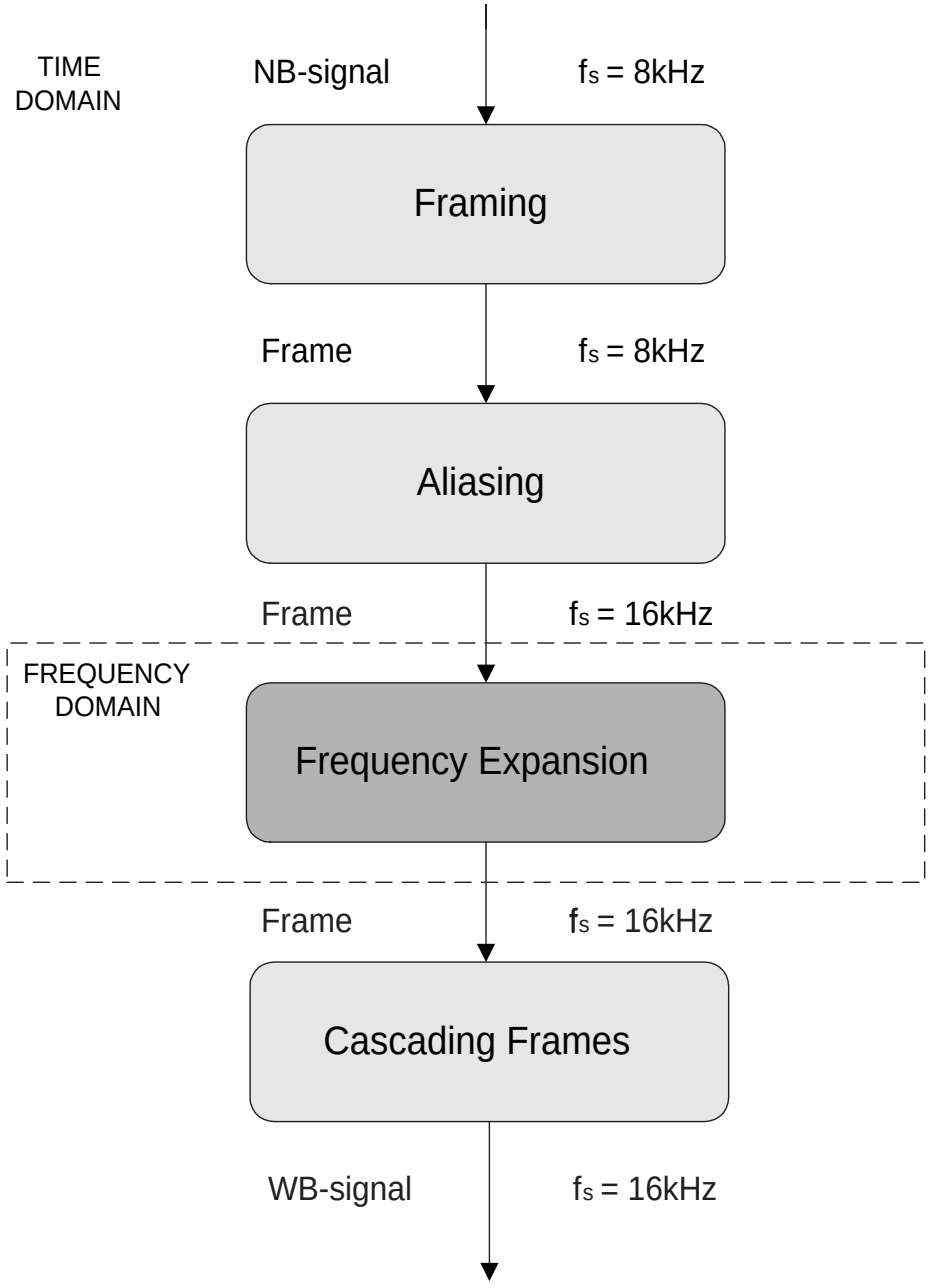

Figure 4.1: Block diagram of the ABE algorithm

due to processing in frames it is possible to make a real time application of the algorithm. The default value for the length of a frame is 20 ms, 160 samples with sampling frequency $f_s = 8kHz$. For extracting a portion of the signal for analysis properly, a window function is used. Each frame is windowed with a Hamming window of length 30 ms, 240 samples. Due to that, every windowed frame overlaps 5 ms with both adjacent frames. Figure 4.2 illustrates the windowing procedure.

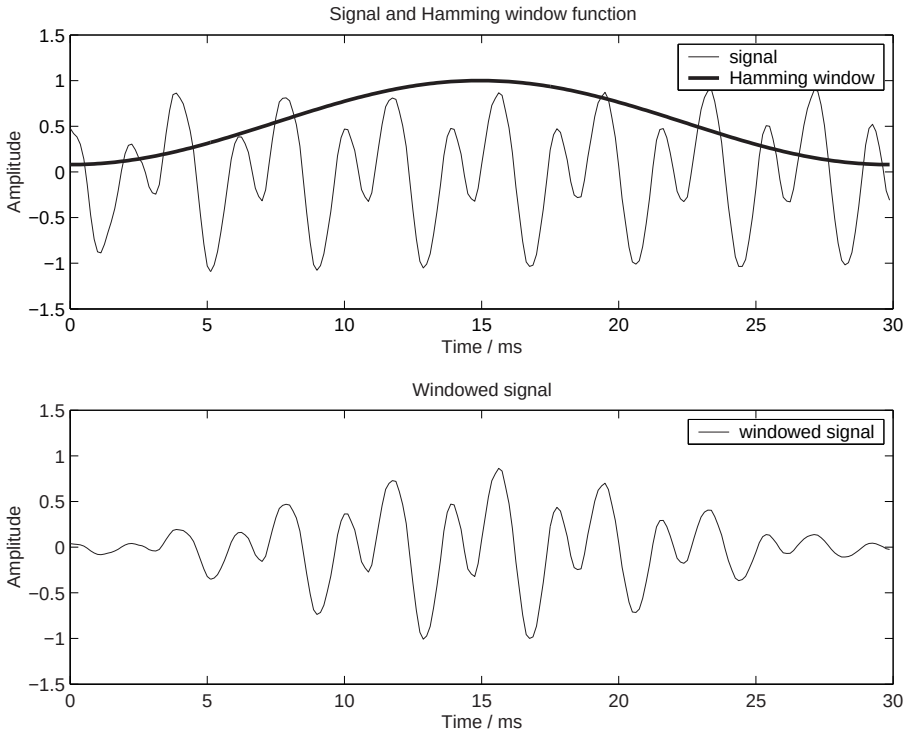

Figure 4.2: Windowing procedure using a Hamming window

The sampling frequency is increased from $8kHz$ to $16kHz$ by up-sampling each narrowband frame by two. The aliased frequencies are preserved in the signal so the spectrum contains the original narrowband spectrum (0–4 kHz) and its mirror image in the highband (4–8 kHz). In practice this up-sampling means that a zero is inserted between every sample and in the frequency domain it corresponds to aliasing.

4.2.3 Frequency Expansion

A block diagram of the frequency expansion is presented in figure 4.3. Sampling frequency of each frame was increased from 8 kHz to 16 kHz through aliasing. Spectral contents of frames are to be modified in the frequency domain in accordance with the classification of sounds into voiced sounds, stop-consonants and sibilants.

Calculation of FFT

The modification of frames is easier to carry out in the frequency domain and that is why a FFT-spectrum is calculated. The length of the FFT is 1024. The FFT spectrum has the original narrowband spectrum in the range of 0-4 kHz and its mirror image from 4 kHz to 8 kHz, like figure 4.4 illustrates.

Modification Functions

After the aliasing and the calculation of the FFT spectrum, a decision whether the frame representing a sibilant or a non-sibilant is needed. Sibilants are fricatives (/s/, /sh/ and /z/) that contain considerably more high frequency components than other phonemes. The separation into sibilants and non-sibilants is made because a spectrum of, for example, a vowel differs considerably from a spectrum of a sound representing a sibilant. For a vowel the spectral envelope is decreasing and for a sibilant it is on the contrary increasing.

The separation of sounds into sibilants and non-sibilants is based on two quotients that are calculated from each frame. Also the duration of the speech sound is taken into account because the duration of a fricative is usually longer than a duration of other consonants. To be even more precise the duration of other fricatives is often less than that of sibilants [AdS01]. Details of quotients and the classification of frames are presented in [Kal02].

As the duration is one of the criteria in deciding whether a speech sound is a sibilant or not, the quotients of the following frame have to be calculated before making that decision. This creates some extra delay which is directly proportional to the size of the frame.

Attenuation Functions for Voiced Sounds and Stop Consonants

Voiced sounds contain more low than high frequency information. Therefore the aliased frequencies have to be attenuated in order to create utterances that sound natural. It was also noticed that if unvoiced stop-consonants (/k/, /p/, /t/) contain too much or a wrong amount of high frequencies, there might be “clicking” or “buzzling” effects in the result. Due to that, spectra of stop-consonants were attenuated as well.

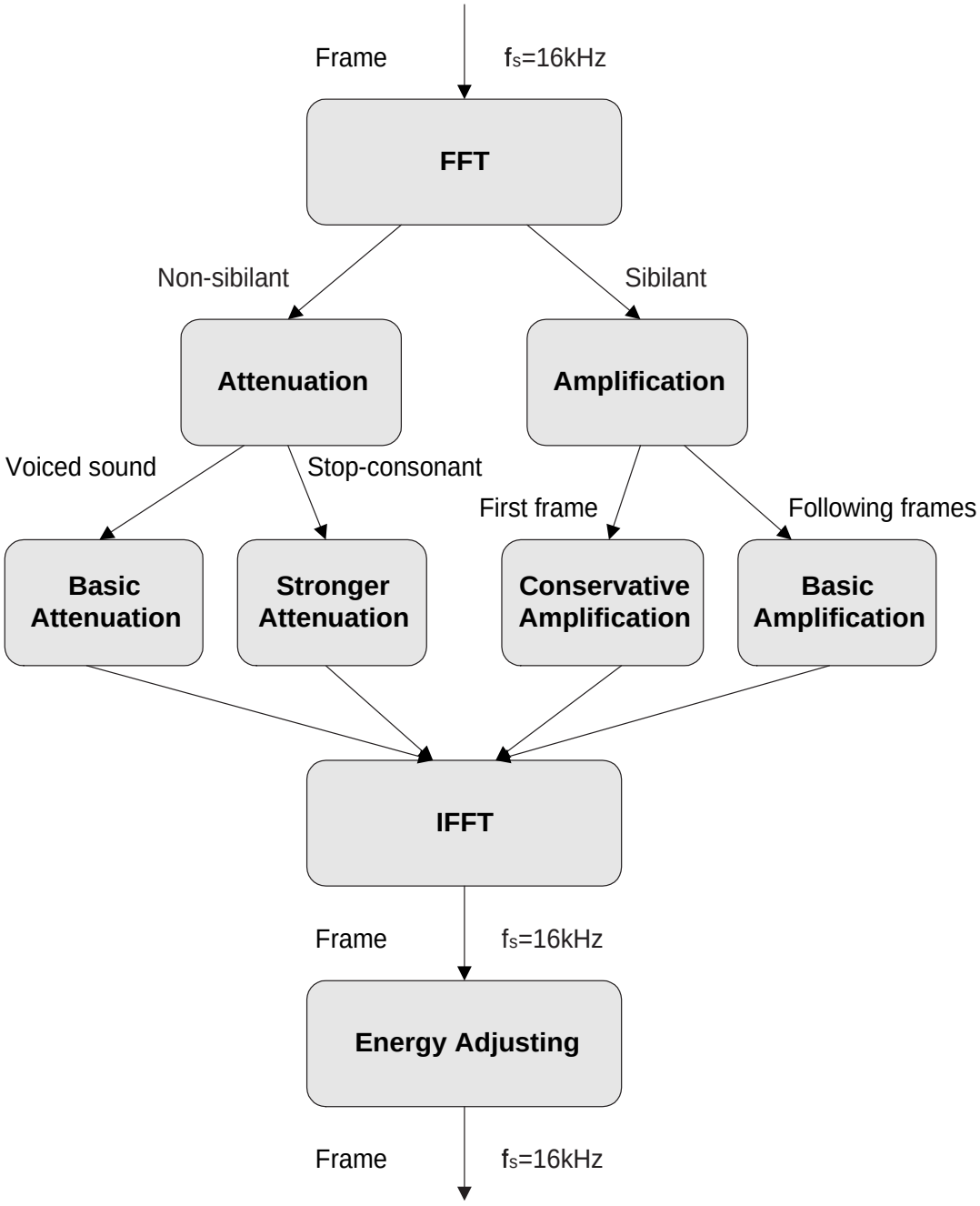

Figure 4.3: Block diagram of the frequency expansion

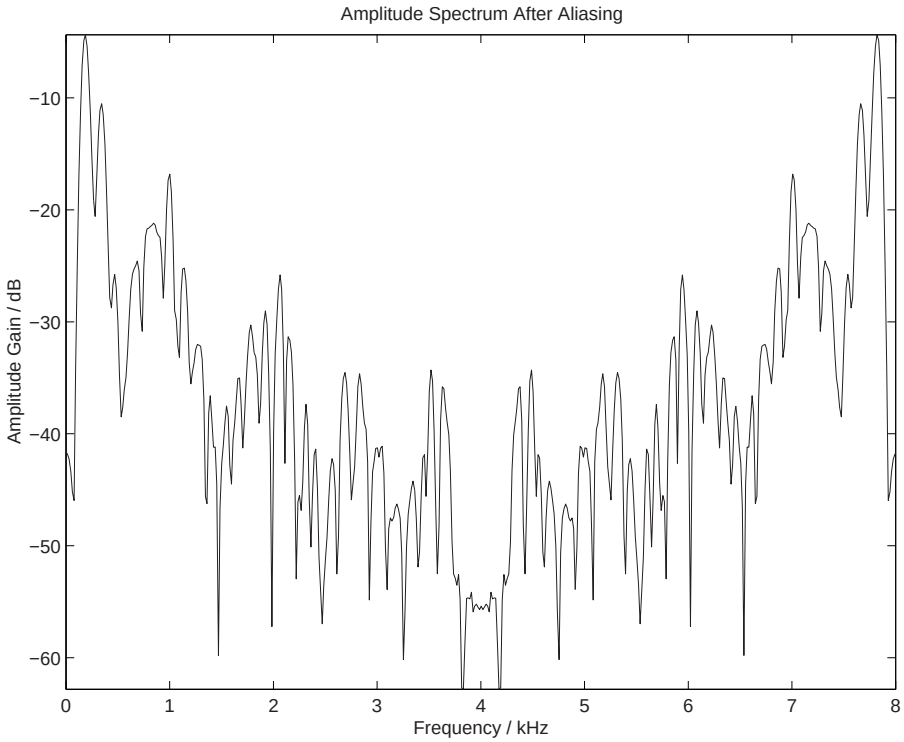

Figure 4.4: An amplitude spectrum of a frame after aliasing

The attenuation function of voiced frames is a descending step-function. Each 1 kHz band is attenuated with a different amount so that the maximum attenuation is proportional to how much the envelope of the amplitude spectrum decreases in the narrowband. The adaptivity of the attenuation function guarantees that no overestimates of the energy in the highband is likely to occur. An example of a spectrum of a voiced frame before and after the attenuation is in figure 4.5.

Stop-consonants are attenuated more than voiced sounds. This guarantees that the higher frequencies do not create “clicking” or “buzzling” effects.

Amplification Function for Sibilants

After the aliasing, the spectrum of a sibilant increases from 0 to 4 kHz and decreases from 4 kHz to 8 kHz. The target is to modify the spectrum so that it will be increasing also in the higher band. To achieve this goal an amplification function is needed.

The narrowband (0–4 kHz) is left untouched and thus the amplification function is zero in that frequency band. The artificially created high band starts from 4 kHz. The amplification function is zero from 4 kHz to 4.8 kHz (20% of the high band). Then it goes up so that the maximum value in the logarithmic scale is proportional to how much the envelope of

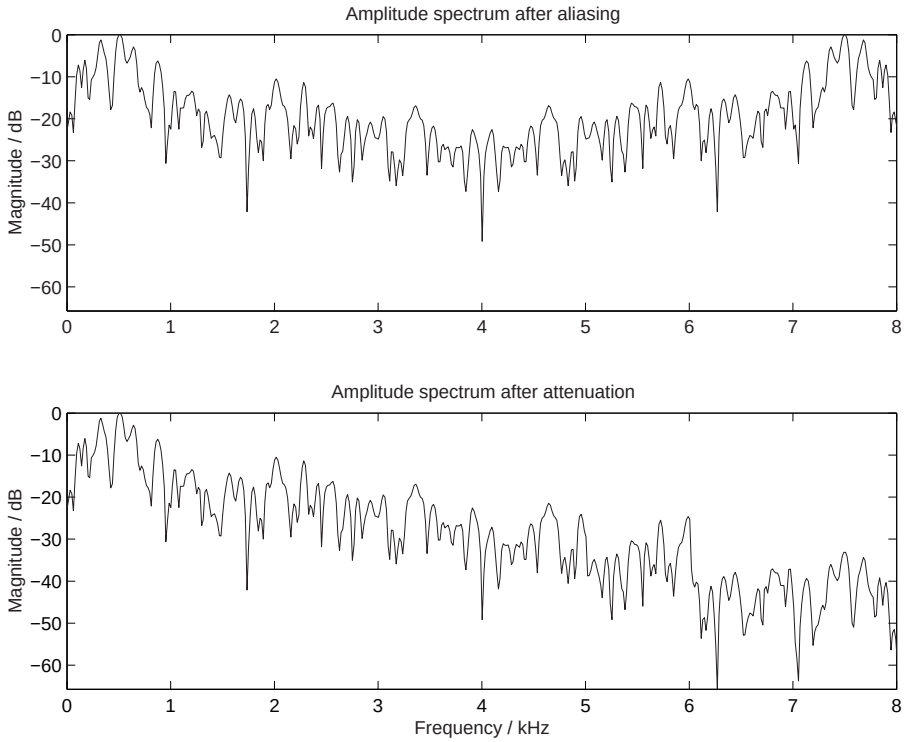

Figure 4.5: An amplitude spectrum after aliasing (top) and the amplitude spectrum after attenuation (bottom)

the amplitude spectrum increases in narrowband. Hence, the amplification function is also adaptive as is the case with the attenuation function. The first frame of each sibilant is processed more conservatively, i.e. amplified less, in order to obtain a smoother transition from non-sibilant to a sibilant.

After the amplification function is determined it is transferred into a linear scale and the aliased spectrum is multiplied by it. Figure 4.6 represents a spectrum of a sibilant before and after the amplification.

Because of the aliasing the harmonic structure of the narrowband appears also in the high band. Spectral peaks of that structure are softened by averaging the highest frequency components.

IFFT and Energy Adjusting

After the processing of the spectrum in the frequency domain, a tranformation into the time domain is needed. This is done with an inverse Fast Fourier Transform, IFFT. An IFFT of

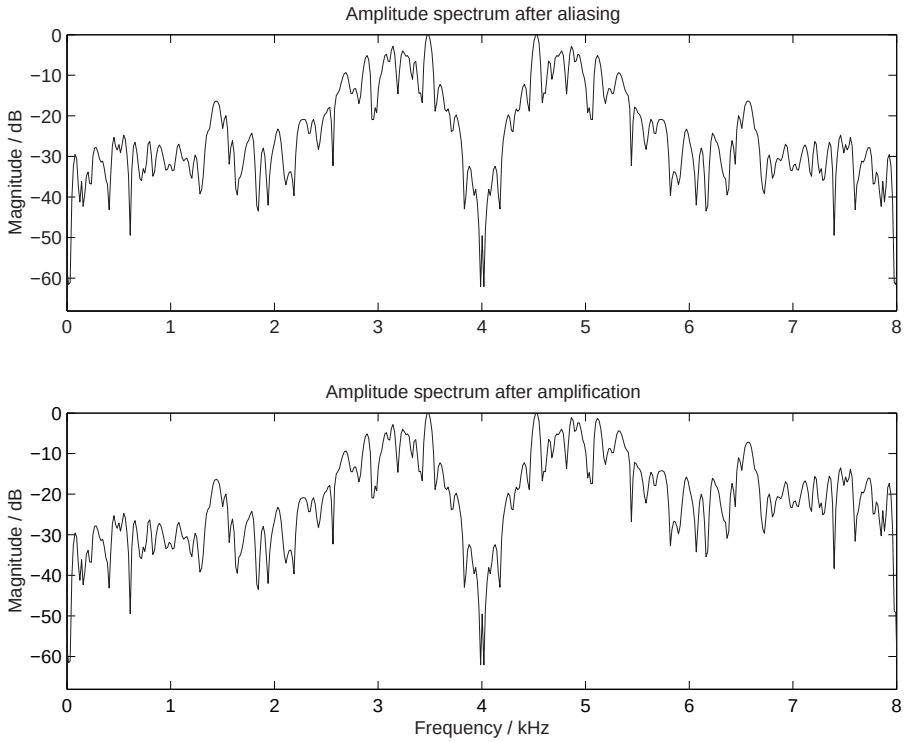

Figure 4.6: An amplitude spectrum after aliasing (top) and the amplitude spectrum after amplification (bottom)

length 1024 is calculated from each frame. From the result 480 first samples (30 ms with sampling frequency $f_s = 16kHz$) form the time domain representation of the frame.

After the frequency expansion, the energy of each frame has changed, because new spectral components are added to the signal. The energy is adjusted by multiplying the expanded signal by 2, as the signal was upsampled by 2.

4.2.4 Unwindowing

All the processed frames are multiplied by an inverse Hamming window in order to compensate the windowing that was done in the computation of the FFT. The length of the inverse window is 30 ms, 480 samples.

4.2.5 Cascading Frames

In order to obtain a continuous signal from the processed frames, the frames are cascaded together. The length of each windowed frame is now 30 ms with sampling frequency $f_s =$

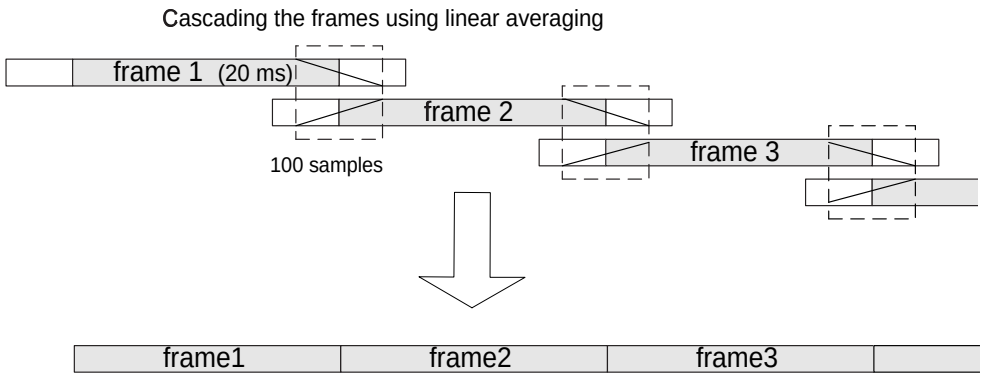

Figure 4.7: Procedure for cascading frames

$16kHz$. The actual frame is the 20 ms part in the middle of the frame. To avoid sudden jumps at the intervals of the frame length, the 50 first and last samples of those actual frames are averaged with samples of the adjacent frames. In the averaging a linear line is used so that, the influence of the following frame increases linearly and at the same time the influence of the current frame decreases linearly. Figure 4.7 illustrates the cascading procedure.

Chapter 5

Results

5.1 Selection of a Suitable Test

After testing and implementing the ABE algorithm with MATLAB, subjective listening tests were arranged at HUT in July 2002 in order to assess the quality of ABE expanded speech signals. The test was a A/B paired comparison test. In this test, signal pairs are presented to the listener. The listener is asked to decide which one is better from every pair, signal A or signal B. From the results a preference score is defined as the percentage of how often listeners choose the concerned signal as the preferred signal.

In the selection of a suitable subjective listening test method the main criterion was that the test should give a clear answer to a question that which one people prefer, conventional narrowband speech or artificially expanded speech. Moreover, it was decided that the test should include different processing chain cases and in order to keep the test simple enough, an A/B paired comparison test was a reasonable choice.

5.2 Test Arrangements

The test was arranged at Helsinki University of Technology, in a listening room which fits the ITU-R BS.116 standard. The software for executing the test was Guinea Pig 2 [HZ99] and Sennheiser HD580 headphones were used in listening. In the first part of the test, 20 naive listeners (12 women and 8 men) listened to mono sample pairs with only one ear (mono test). They were asked to decide which one is more agreeable (in Finnish “miellyttävä”) and which one he or she would rather listen to (in Finnish “kumpaa kuuntelisit mieluummin”). All the samples were in Finnish and also all the subjects were

native Finnish speaking. The sentences used in the test are in table 5.1. The second part of the test was a smaller separate test. In it, the same subjects listened to a portion of the mono sample pairs with both ears (stereo test).

Test Samples

The test samples consisted of 16 different Finnish sentences, 8 spoken by female speakers and 8 by male speakers plus one zero sample. The sentences and speakers are given in table 5.1. The processing of 8 first sentences (P1) corresponded to the processing of calls originated from landline telephone whereas the processing of the 8 last sentences (P2) was similar to calls originated from a mobile terminal. The zero sample was processed in both ways. Inside these two main processing blocks, P1 and P2, there were three different processing chains. The first one (A) corresponded to how the conventional narrowband speech is processed in telecommunication network. The second (B) refers to a potential case where ABE is implemented in a terminal and the third processing chain (C) refers to a situation when ABE is implemented on the network side. The exact processing chains are:

```
P1:
A: IRS - PCM codec - EFR - upsampling to 16 kHz
B: IRS - PCM codec - EFR - ABE
C: IRS - PCM codec - ABE - AMR-WB

P2:
A: GSM1 - EFR - EFR - upsampling to 16 kHz
B: GSM1 - EFR - EFR - ABE
C: GSM1 - EFR - ABE - AMR-WB
```

The abbreviations are:

IRS refers to the ITU-T modified IRS weighting filter that models the send characteristics of a common landline telephone, see figure 5.1

PCM codec means PCM transmit line filter characteristics (CCITT G.712), which is a bandpass filter used together with analog-to-digital conversion, see figure 5.2

EFR EFR speech codec

GSM1 stands for ITU-T GSM mobile station input characteristics, see figure 5.3

ABE is the implemented artificial bandwidth expansion algorithm

AMR-WB AMR wideband speech codec (mode 12.65 kbit/s)

A is the original narrowband signal

B refers to ABE implemented in a terminal

C refers to ABE implemented on the network side

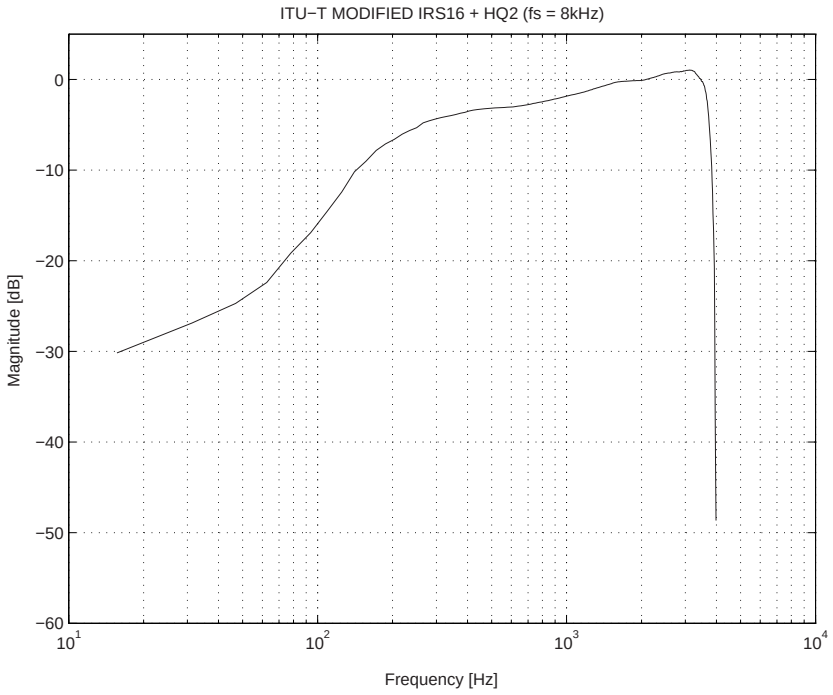

Figure 5.1: Frequency response of IRS16 filter (ITU-T)

In the first part of the test (mono listening) all the sample pairs AB, AC and BC, were compared. In the second part (stereo listening) only AB pairs were compared. In addition to different processing chains, different background noise situations were used as well:

Noise situations for P1:

```
N1:  office noise SNR > 30 dB
N2:  babble noise SNR = 10 dB
```

Noise situations for P2:

```
N1:  office noise SNR > 30 dB
N2:  babble noise SNR = 10 dB
N3:  car noise SNR = 10 dB
```

Background noise was added to speech samples before any processing was done. So for example in the case of babble noise samples, the SNR was 10 dB before any processing was done and final versions of them may have a slightly different SNR.

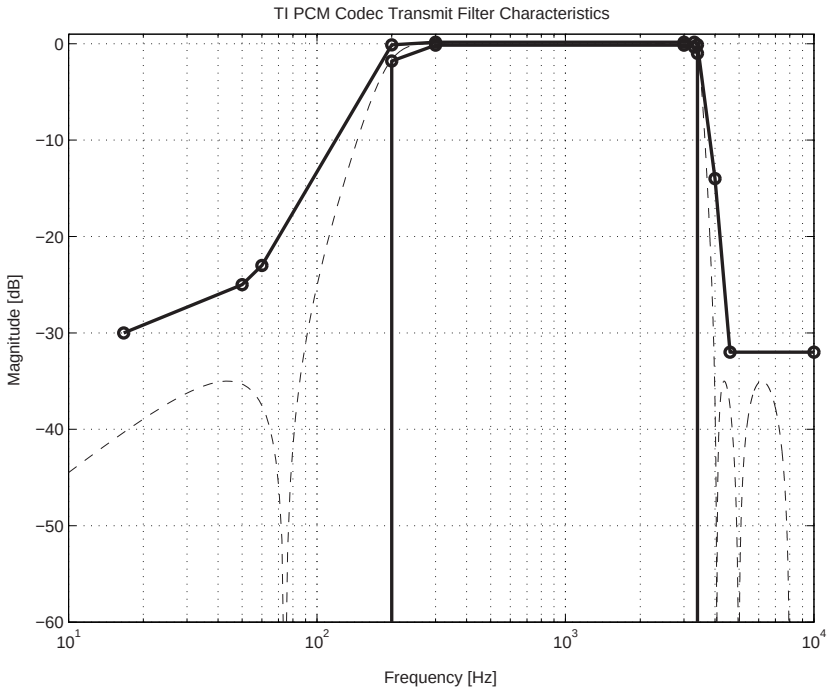

Figure 5.2: Frequency response of PCM filter [Tex89]

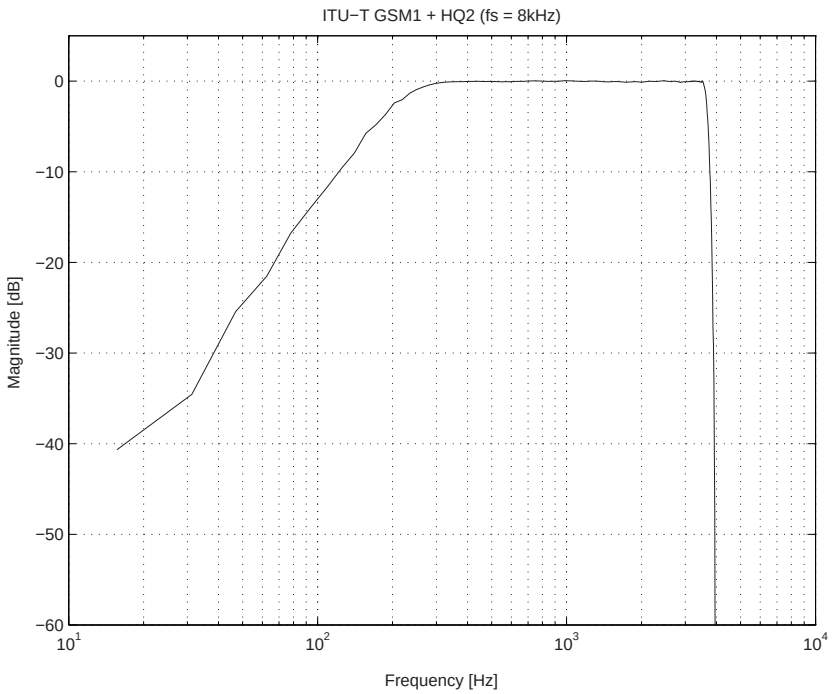

Figure 5.3: Frequency response of GSM1 filter (ITU-T)

Number	Speaker	Sentence
s1	female 1	Kenkien tulisi olla väljät.
s2	female 2	Varas sieppasi timanttisormuksen.
s3	female 3	Minä olen opiskelija.
s4	female 4	En tiedä mikä tämä on.
s5	male 1	Asiantuntijat olivat yksimielisiä.
s6	male 2	Aurinko nousee idästä.
s7	male 3	Kissä näkee hyvin pimeässä.
s8	male 4	Koe onnistui yli odotusten.
s9		*zero sample*
s10	female 1	Kaikki odottavat talouden elpymistä.
s11	female 3	Käärme nieli saaliinsa.
s12	female 4	Valitettavasti en voi auttaa.
s13	female 5	Juna lähtee hetken kuluttua.
s14	male 1	Jätä viesti puhelinvastaajaan.
s15	male 5	Poliisi tutkii asiaa.
s16	male 6	Ennakkosuosikki voitti kilpailun.
s17	male 7	Olutta läikkyi lattialle.
s18		*zero sample*

Table 5.1: Sentences used in the listening test

The ABE algorithm which is described in chapter 4 was used in office noise and car noise situations. In babble noise situations, a more conservative version was used because the babble noise situation is more demanding for the ABE algorithm. In the conservative version, all the frames were processed as stop-consonants, which means that the attenuation of the high band is very deep. The number of unpleasant “click sounds” was decreased but on the other hand the difference between the narrowband signal and the expanded signal is rather small.

The first part of the test comprised all the possible processing and noise cases but the second part had only P1 processing with office and babble noise cases. In total, there were 45 AB-test pairs, 45 AC-test pairs and 45 BC-test pairs in the mono part of the test and 36 AB-pairs in the stereo part of the test.

5.3 Results

Confidence intervals of the results were calculated using binomial distribution, because the number of events was too small for normal distribution. The formula for upper limit (U.L.) is:

$$U.L. = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\phi(1-\phi)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n} \tag{5.1}$$

and for lower limit (L.L.):

$$L.L. = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\phi(1-\phi)}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \tag{5.2}$$

where n is the sample size, $\hat{p}$ is the proportional defective and $z_{\alpha/2}$ is the test statistic (the value that is exceeded with probability $\alpha/2$) [Han].

Preference scores for narrowband and ABE-terminal signals of mono listening are in figure 5.4. Equivalent preference scores for narrowband – ABE-network and ABE-terminal – ABE-network pairs are in figures 5.5 and 5.6. More detailed results are presented in Appendix A. Sample pairs that contained only background noise were analyzed separately. Preference scores for them can be found in figures 5.7–5.9.

Results of the stereo listening part of the test are in figures 5.10 and 5.11. The former contains preference scores for speech samples and the latter for background noise samples.

Preference scores for all the sentences used in the listening test are presented in figure 5.12. The sentences and the speakers can be seen in table 5.1. Figure 5.13 represents how the preference scores for ABE samples of all the subjects have been distributed in office noise cases.

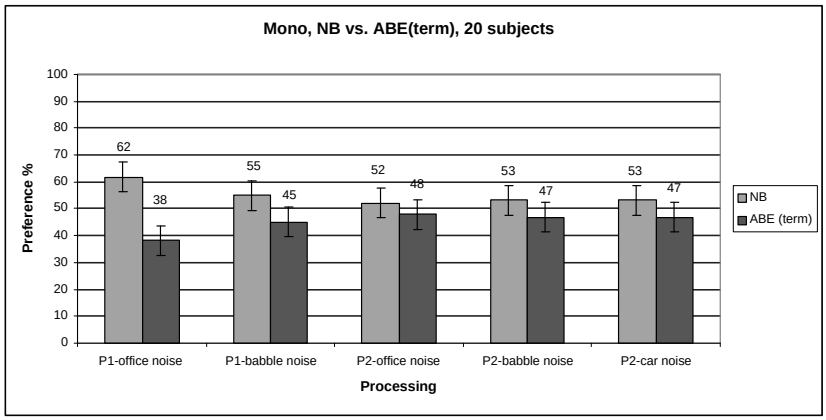

Figure 5.4: Preference scores with confidence intervals calculated from the answers of all the subjects. Mono listening, processings P1 (calls originated from landline telephone) and P2 (calls originated from a mobile terminal), narrowband vs. ABE-terminal, office, babble and car noise cases

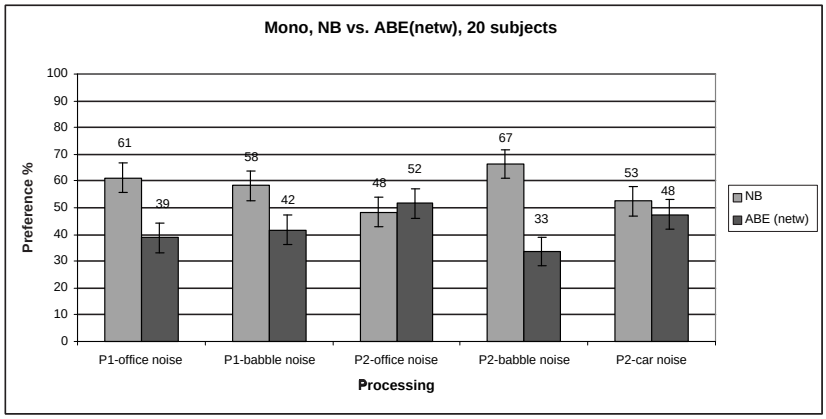

Figure 5.5: Preference scores with confidence intervals calculated from the answers of all the subjects. Mono listening, processings P1 (calls originated from landline telephone) and P2 (calls originated from a mobile terminal), narrowband vs. ABE-network, office, babble and car noise cases

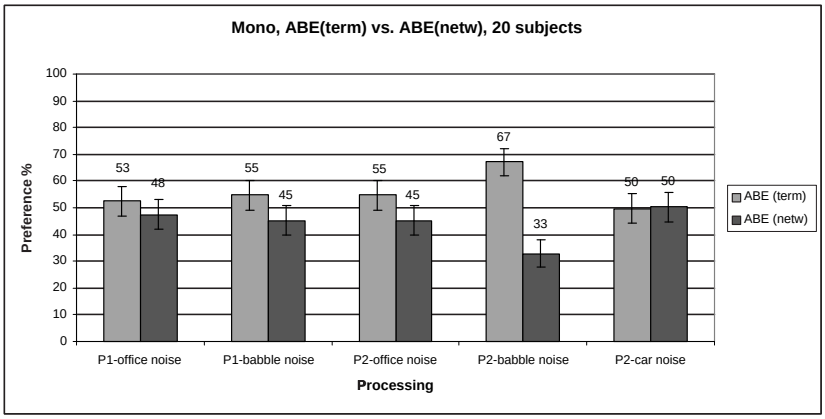

Figure 5.6: Preference scores with confidence intervals calculated from the answers of all the subjects. Mono listening, processings P1 (calls originated from landline telephone) and P2 (calls originated from a mobile terminal), ABE-terminal vs. ABE-network, office, babble and car noise cases

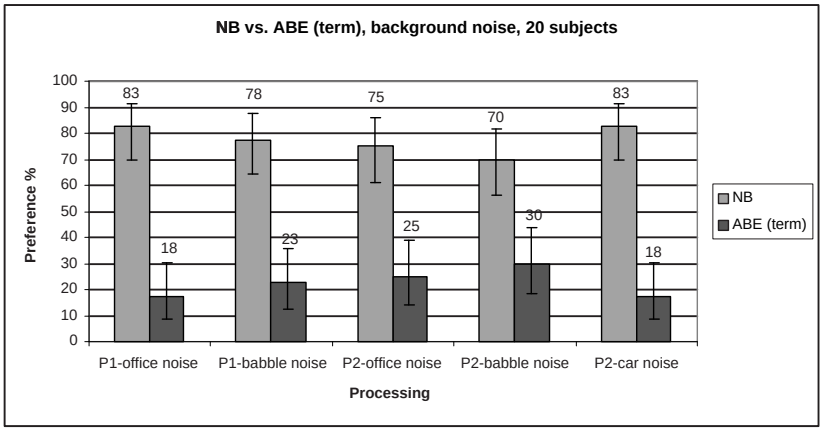

Figure 5.7: Preference scores with confidence intervals for background noise samples, narrowband vs. ABE-terminal

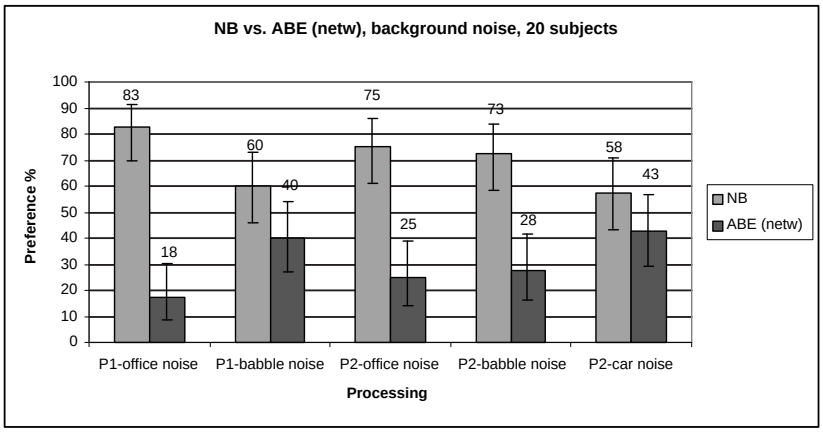

Figure 5.8: Preference scores with confidence intervals for background noise samples, narrowband vs. ABE-network

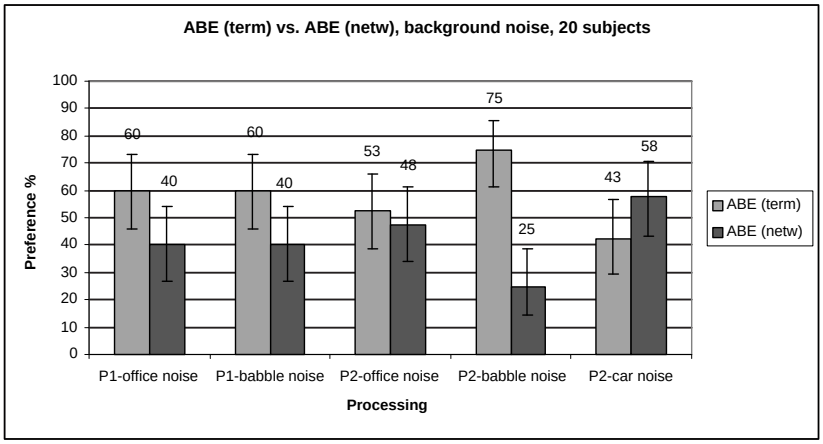

Figure 5.9: Preference scores with confidence intervals for background noise samples, ABE-terminal vs. ABE-network

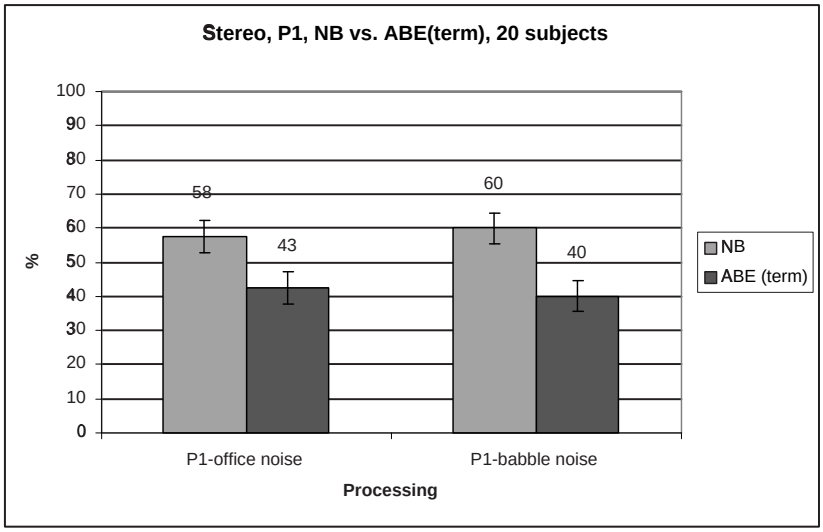

Figure 5.10: Preference scores with confidence intervals calculated from the answers of all the subjects. Stereo listening, processing P1 (calls originated from landline telephone), narrowband vs. ABE-terminal, office and babble noise cases

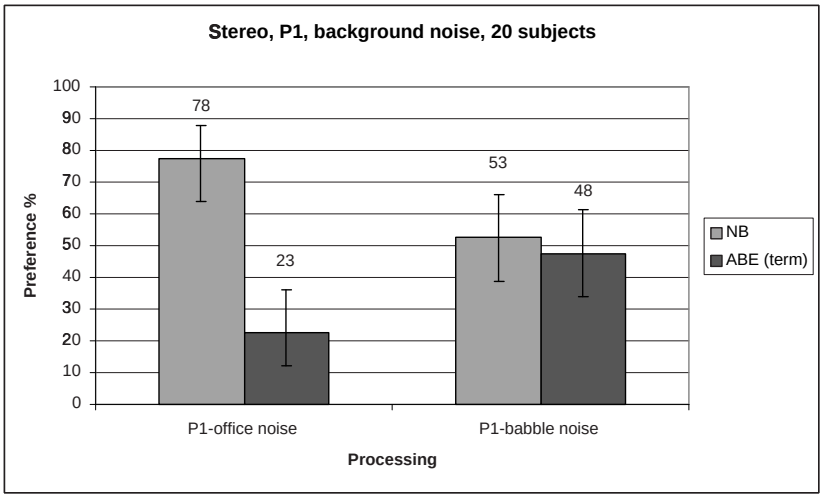

Figure 5.11: Preference scores with confidence intervals for stereo background noise samples, narrowband vs. ABE-terminal

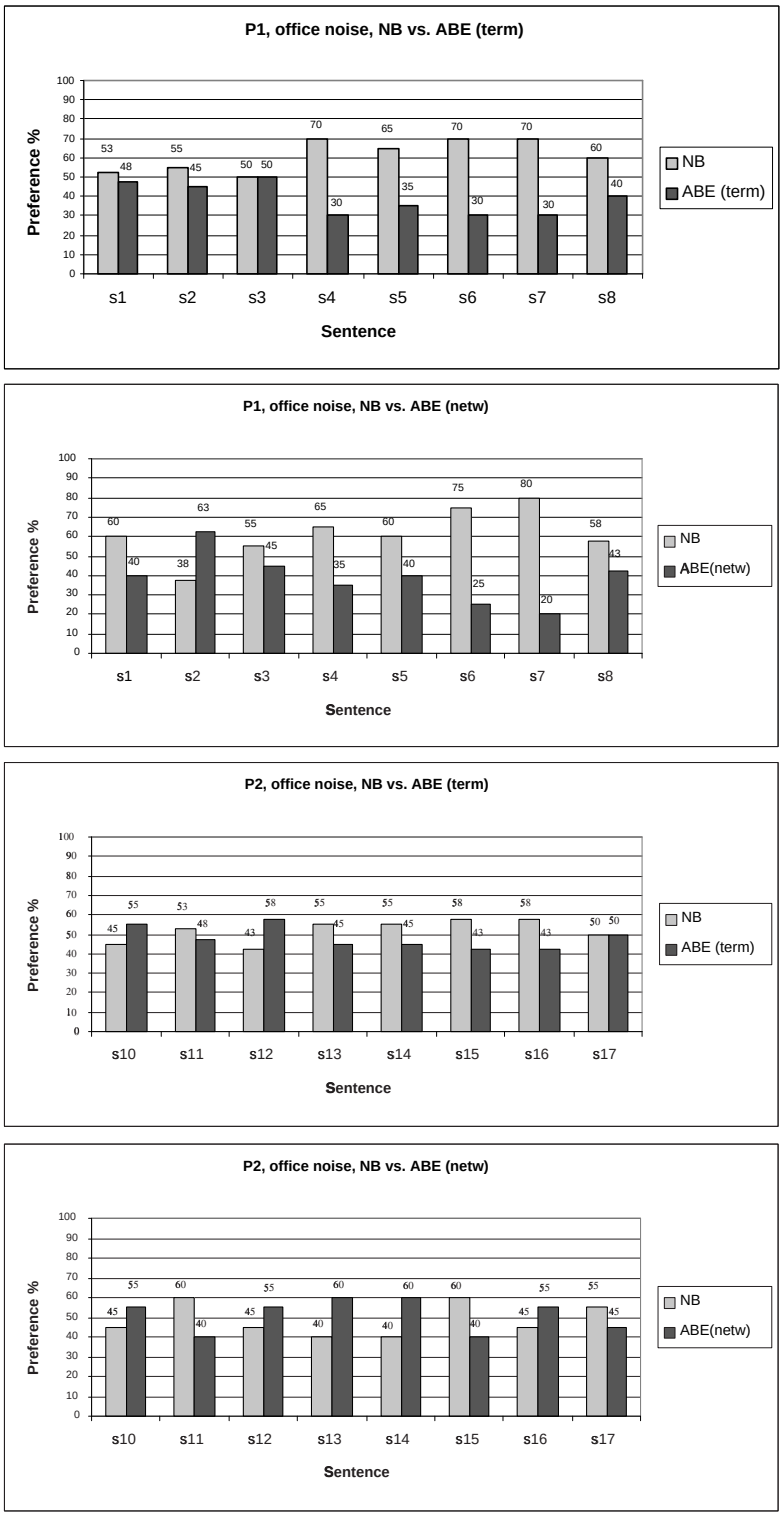

Figure 5.12: Preference scores for sentences used in the listening test

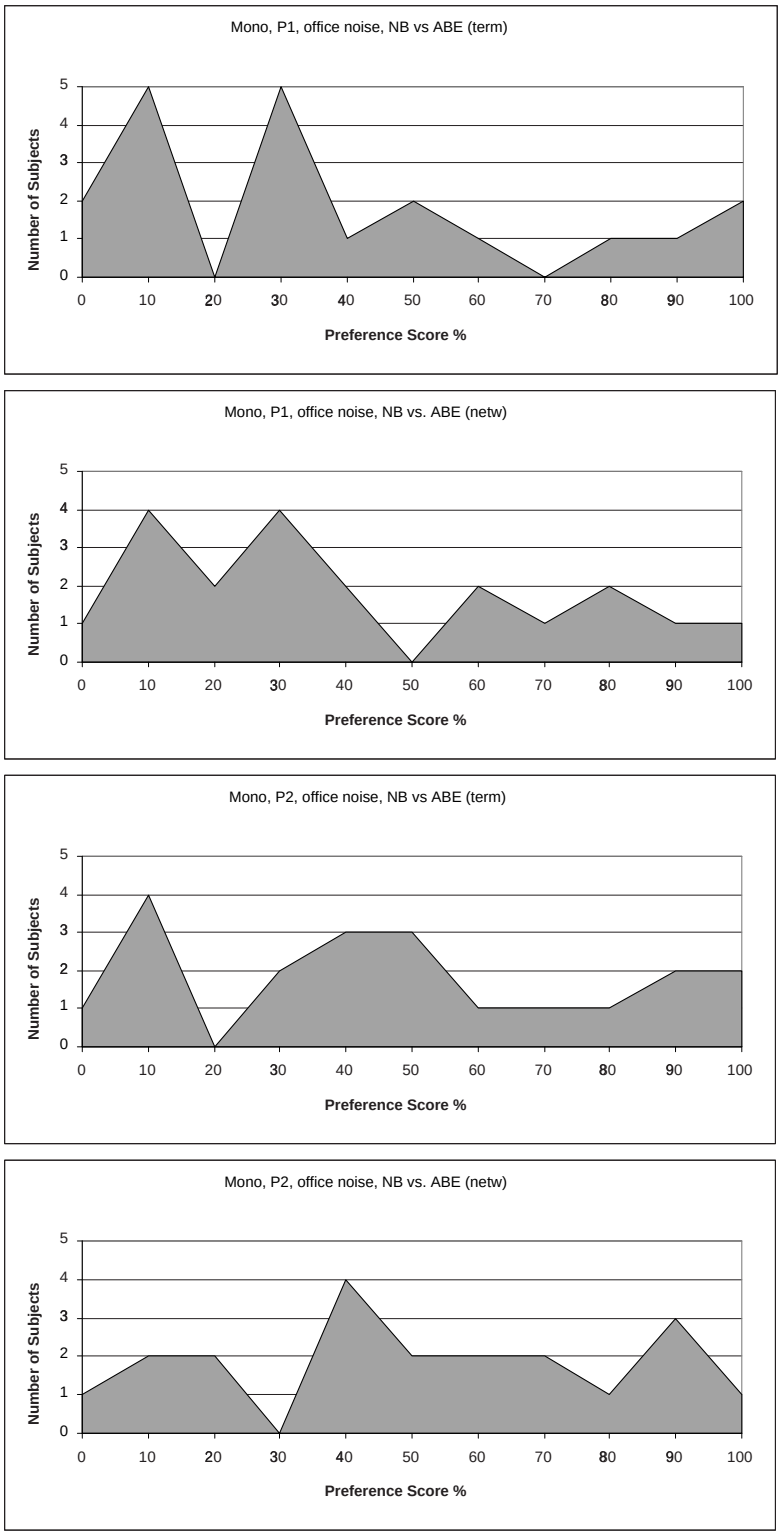

Figure 5.13: Preference scores for ABE-expanded samples as a function of the number of subjects in office noise cases

Chapter 6

Conclusions

The ABE algorithm is based on simple aliasing. The input signal is processed in short frames. By calculating the FFT and IFFT it is easy to move to the frequency domain and back to the time domain. The frequency band is expanded by creating a mirror image from 0-4 kHz to 4-8 kHz. The mirror image is attenuated or amplified in the frequency domain and finally after the IFFT frames are cascaded.

The ABE algorithm works at its best with clean speech samples but also samples with flat noise are expanded properly with it. Babble noise situation is more demanding. Some frames containing “clinks” caused by dishes or a rustle of papers are often processed as if they were sibilants and the result contains very annoying “clicking” sounds.

Although the default frame length is quite long, the result of the ABE algorithm sounds the same in most cases regardless of in what exact places the edges of the frames happen to be. In that way the algorithm is very robust.

One criterion for the sibilant detection is the duration of it. There has to be at least two adjacent frames that are considered sibilants before the final decision of which processing method will be used, is made. So if it seems that a current frame is a sibilant, the algorithm looks forward one frame and that causes extra delay to the algorithm. The delay is directly proportional to the size of a frame which has the default value of 20 ms. To make that delay shorter, the size of a frame should be smaller. It is possible to shorten it and it doesn’t seem to spoil the output of the algorithm. Another way of shortening the delay is to process not the first but the second and the following frames that the sibilant detection algorithm has detected as sibilants, but the quality of expanded speech might not be as good in that case.

As a result of the aliasing, there is a gap in the spectrum around 4 kHz. It was tested that the gap does not have a significant effect on the quality of artificially expanded speech. Human

hearing is not as sensible to sudden gaps in the spectrum as it is to sudden peaks.

6.1 Remarks on Results

From the results of the listening test introduced in chapter 5 it can be concluded that the overall result is that listeners gave a slightly better score for narrowband signals compared to ABE-signals. There are only small differences between different background noise and processing chain cases, so the ABE algorithm seems to work as well in all the cases. The difference between ABE-terminal and ABE-network samples was only just audible, but the results show that ABE-terminal versions of samples have better preference scores. The stereo test gave very similar results as the mono test, preference score for ABE-terminal samples is approximately 40–45%. When the confidence intervals are taken into account, the difference between preference scores of narrowband signals and ABE signals gets even minor.

The results in figure 5.12 show that some sentences expanded with the ABE algorithm got better preference scores than others. On the other hand, none of the sentences got very bad score, which supports the conclusion that the algorithm is robust.

There are many reasons for why narrowband signals were ranked better than ABE samples. Firstly and the most importantly, people are used to listening to narrowband signals. They sound familiar and are a safe choice in a new test situation. Secondly, some listeners listened to the background noise more than others. The preference scores for background noise samples in figures 5.7, 5.8, 5.9 and 5.11 show that subjects did not like how the algorithm affected them. The ABE algorithm added higher frequency components to the background noise which made it more distracting in some listeners' opinion. These higher frequency components can be heard also in samples that contain both speech and noise which affected the results of the listening test.

The preference scores that every listener gave in each test situation are presented in Appendix A. From the results it can be concluded that the opinions of the subjects vary a lot. Others value clear phonemes whereas others prefer a soft general impression. The subjects that participated the listening test gave oral comments on the sound samples. Some of them can be seen in Appendix B. All the subjects said that they heard a clear difference between most of the sample pairs. Nobody commented that any of the samples would have been unnatural or unreal and that is already a positive result from the test. A conclusion can be made that there are no bad artifacts in the ABE sound samples. All the sound samples were considered real and eventually the decision that every subject made, was based on his or her

own experience and opinion on what kind of characteristics of speech he or she prefers.

Figure 5.13 presents how the preference scores for ABE samples of all the subjects have been distributed in office noise cases. It can be seen that different subjects have very different opinions. Others give a very low preference score to ABE samples and others very high. This confirms the conclusion that the decision on which one is more agreeable, the narrowband or the ABE version, is dependent on the listener and what kind of characteristics of speech he or she appreciates.

6.2 Future Work

The ABE algorithm is based on simple aliasing, which is at the same time an advantage and a disadvantage. On one hand it makes the algorithm simple and safe but on the other hand there is a limit for how much further the algorithm can be developed. In this case, some improvements could still be possible and from the listening test results a conclusion can be drawn that they are also needed. A few ideas for possible future work are discussed next.

The ABE algorithm could be more conservative in some degree. That was at least the impression that was got from the comments of the subjects of the listening test. If the difference between the old narrowband signal and the expanded version of it was somewhat minor, it could be easier for ordinary people to accept it. Moreover with more conservative algorithm, possible artifacts would not be so distracting.

When compared to real wideband speech, ABE samples sound more metallic. Also many of the subjects that participated the listening test had noticed it and said that the “hardness” and the “sharpness” of the expanded speech samples was disturbing. A paired comparison between the narrowband version and the ABE version is probably not the best way to examine the quality of ABE samples. This is because some samples differ so much from each other that the comparison is easily made only on the grounds of the tone of voice. A subject might give a better score for narrowband sample only because the voice tone of it sounds familiar or pleases him or her more.

The number of incorrect sibilant detections could be reduced by activating the detection only if the frame contains speech. This could be implemented using the voice activity detector (VAD). That way annoying “clicks” and other artifacts would diminish and specially speech samples with babble noise would sound much better after artificial expansion.

To summarize, the ABE algorithm presented within this thesis has given promising results. Some improvements are still needed but the overall framework is already working. In the future when the wideband transmission will become more general, both narrowband and

wideband signals will be transmitted in the same network. Therefore, there is a need for a robust artificial bandwidth expansion method that would facilitate the compatibility of systems transmitting signals of different frequency bandwidth.

Bibliography

[AdS01] A. M. A. Ali and J. V. der Spiegel. Acoustic-phonetic Features for the Automatic Classification of Fricatives. *Journal of Acoustical Society of America*, 109(5):2217–2235, May 2001.

[AHW95] C. Avendano, H. Hermansky, and E. A. Wan. Beyond Nyquist: Towards the Recovery of Broad-bandwidth Speech from Narrow-bandwidth Speech. In *Proceedings of European Conference on Speech Communication and Technology, Madrid*, 1995.

[CGMS01] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter. Speech Enhancement via Frequency Bandwidth Extension Using Line Spectral Frequencies. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 665–668, 2001.

[CH96] C.-F. Chan and W.-K. Hui. Wideband Re-synthesis of Narrowband CELP-coded Speech Using Multiband Excitation Model. In *Proceedings of International Conference on Spoken Language Processing*, pages 322–325, 1996.

[CH97] C.-F. Chan and W.-K. Hui. Quality Enhancement of Narrowband CELP-coded Speech via Wideband Harmonic Re-synthesis. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 1187–1190, 1997.

[COM94] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein. Statistical Recovery of Wideband Speech from Narrowband Speech. In *IEEE Transactions Speech and Audio Processing*, volume 2, pages 544–548, Oct. 1994.

[Dim91] S. Dimolitsas. Subjective Quality Quantification of Digital Voice Communication Systems. In *IEEE Proceedings-1*, volume 138, pages 585–595, December 1991.

[EH99] J. Epps and W. H. Holmes. A New Technique for Wideband Enhancement of Coded Narrowband Speech. In *IEEE Workshop on Speech Coding, Porvoo, Finland*, pages 174–176, 1999.

[EK99] N. Enbom and W. B. Kleijn. Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients. In *IEEE Workshop on Speech Coding, Porvoo, Finland*, pages 171–173, 1999.

[Han] Engineering Statistics Handbook. Section 7.2.4.1, Confidence intervals. http://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm (Oct 7, 2002).

[HNK02] M. Hosoki, T. Nagai, and A. Kurematsu. Speech Signal Bandwidth Extension and Noise Removal Using Subband HMM. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida*, pages 245–248, 2002.

[HZ99] J. Hynninen and N. Zacharov. Guineapig - A generic subjective test system for multichannel audio. In *106th Audio Engineering Society Convention, Munich, Germany*, May 1999.

[JV00] P. Jax and P. Vary. Wideband Extension of Telephone Speech Using a Hidden Markov Model. In *IEEE Workshop on Speech Coding, Delavan, USA*, pages 130–135, 2000.

[JV02] P. Jax and P. Vary. An Upper Bound on the Quality of Artificial Bandwidth Extension of Narrowband Speech Signals. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida*, pages 237–240, 2002.

[Kal02] L. Kallio. Artificial Bandwidth Expansion of Telephone Speech with Special Emphasis on Processing of Fricatives. Project Report, Halsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2002.

[Kar00] M. Karjalainen. *Kommunikaatioakustiikka*. Teknillinen korkeakoulu, 2000.

[KN88] N. Kitawaki and H. Nagabuchi. Quality Assessment of Speech Coding and Synthesis Systems. *IEEE Communications Magazine*, pages 36–44, October 1988.

[Käy01] K. Käyhkö. A Robust Wideband Enhancement for Narrowband Speech Signal. Research Report, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2001.

[Mah99] J. Mahkonen. Äänen laadun parantaminen puheensiirrossa keinotekoisella taajuuskaistan laajennuksella. Master's thesis, Teknillinen korkeakoulu, 1999.

[NAK00] M. Nilsson, S. V. Andersen, and B. Kleijn. On the Mutual Information Between Frequency Bands in Speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1327–1330, June 2000.

[NGAK02] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn. Gaussian Mixture Model Based on Mutual Information Estimation Between Frequency Bands in Speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida*, pages 525–528, 2002.

[P.896] ITU-T Recommendation P.800. Series P: Telephone Transmission Quality, Methods for Objective and Subjective Assessment of Quality, August 1996.

[PK00] K.-Y. Park and H. S. Kim. Narrowband to Wideband Conversion of Speech Using GMM Based Transformation. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 1843–1846, 2000.

[Ric73] D. L. Richards. *Telecommunication by Speech*. Wiley, New York, 1973.

[Tex89] Texas Instruments. *SCTS030E*, August 1989. Revised October 1996.

[Too82] F. E. Toole. Listening Tests - Turning Opinion into Fact. *Audio Engineering Society*, 30(6):431–4445, June 1982.

[Yas94] H. Yasukawa. Quality Enhancement of Band Limited Speech by Filtering and Multirate Techniques. In *Proceedings of International Conference on Spoken Language Processing*, pages 1607–1610, 1994.

[Yas96] H. Yasukawa. Signal Restoration of Broad Band Speech Using Nonlinear Processing. In *Proceedings of European Signal Processing Conference*, pages 987–990, 1996.

[Yas98] H. Yasukawa. Wideband Speech Recovery from Bandlimited Speech in Telephone Communications. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, volume 4, pages 202–204, 1998.

Appendix A

Listening Test Results

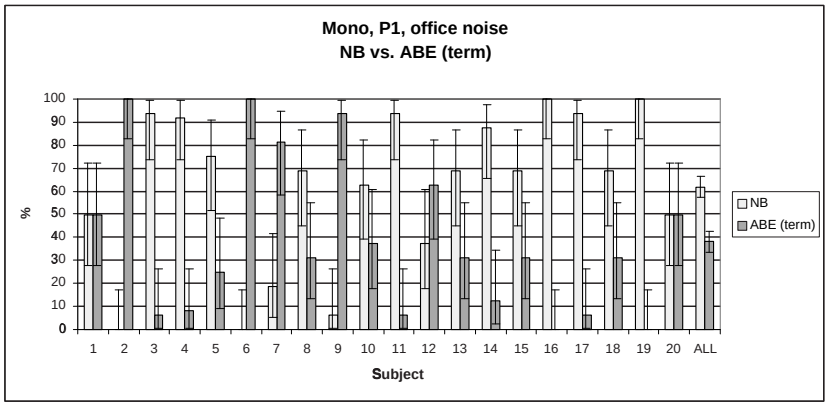

Figure A.1: Preference scores with confidence intervals of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), narrowband vs. ABE-terminal, office noise

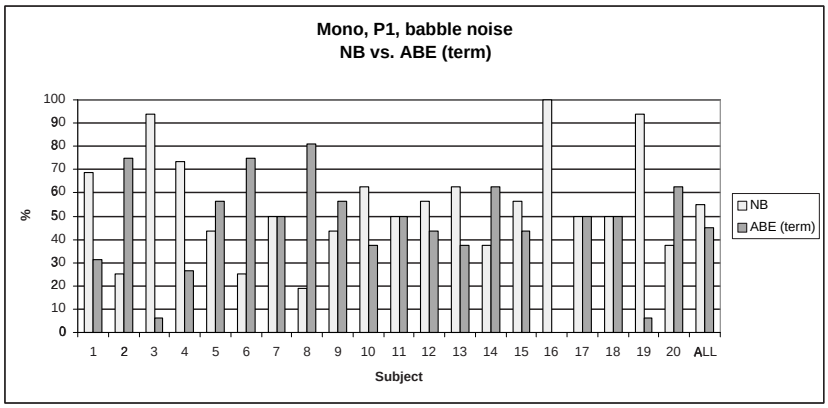

Figure A.2: Preference scores of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), narrowband vs. ABE-terminal, babble noise

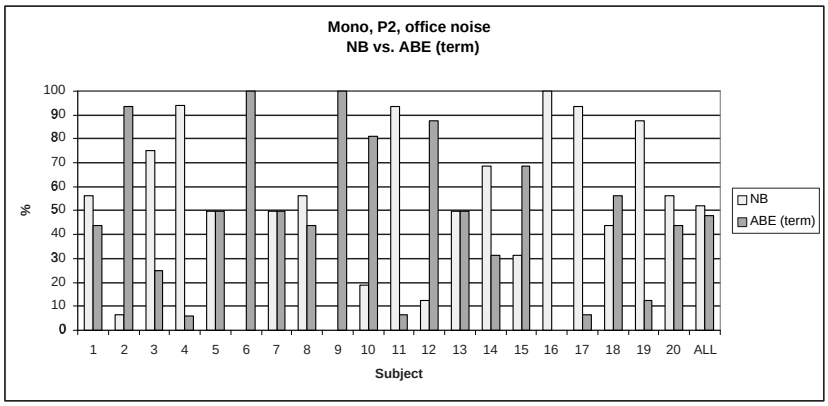

Figure A.3: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-terminal, office noise

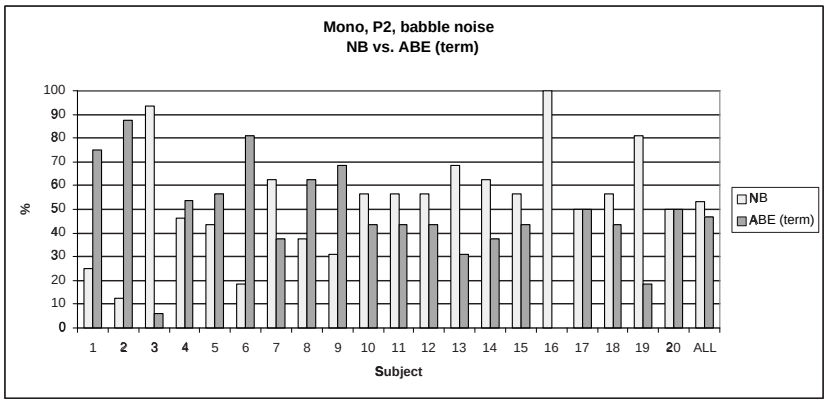

Figure A.4: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-terminal, babble noise

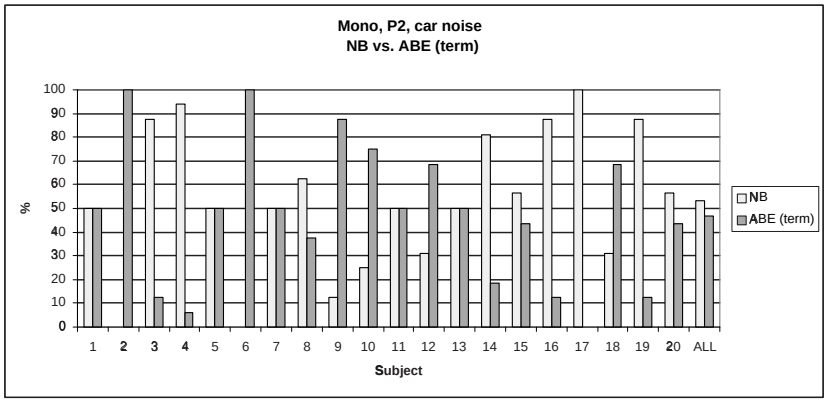

Figure A.5: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-terminal, car noise

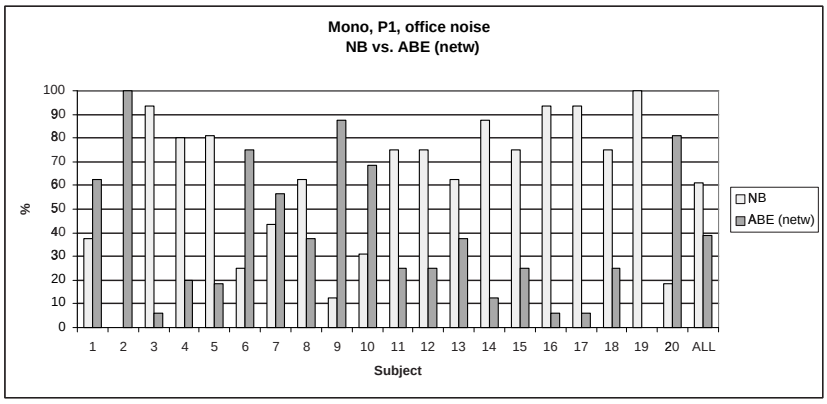

Figure A.6: Preference scores of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), narrowband vs. ABE-network, office noise

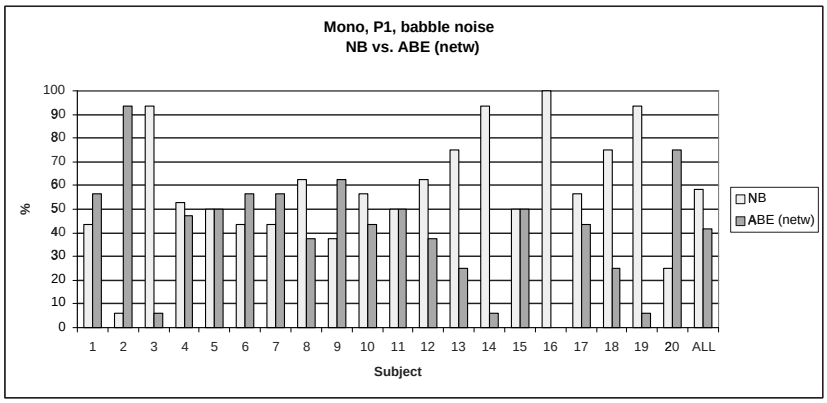

Figure A.7: Preference scores of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), narrowband vs. ABE-network, babble noise

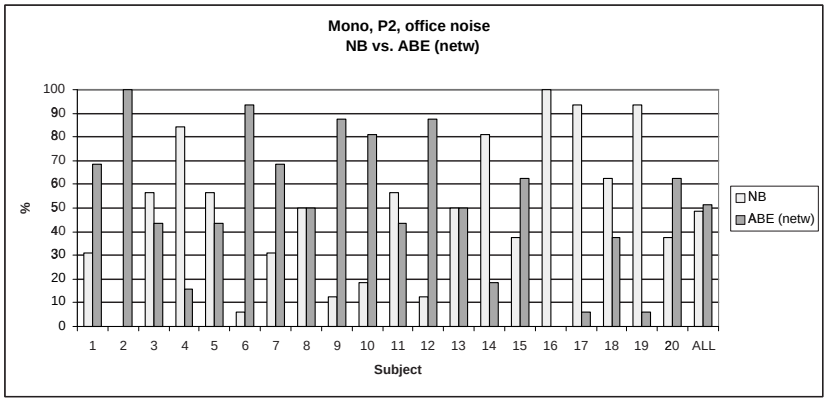

Figure A.8: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-network, office noise

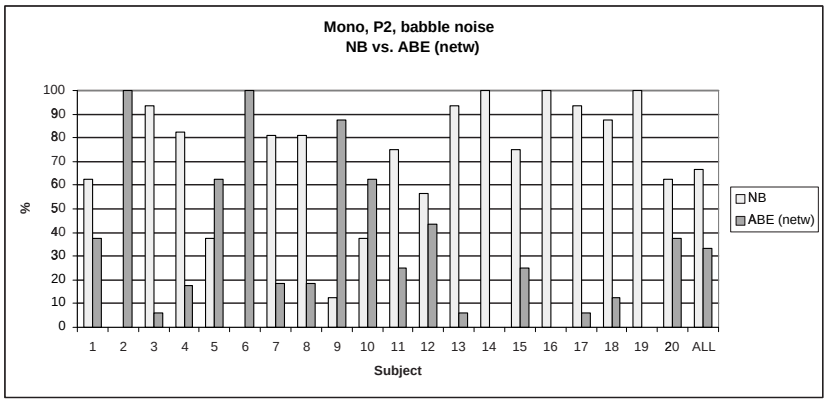

Figure A.9: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-network, babble noise

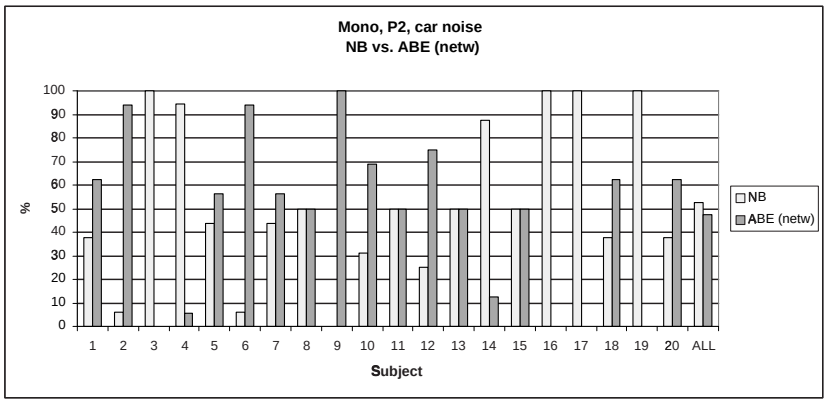

Figure A.10: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), narrowband vs. ABE-network, car noise

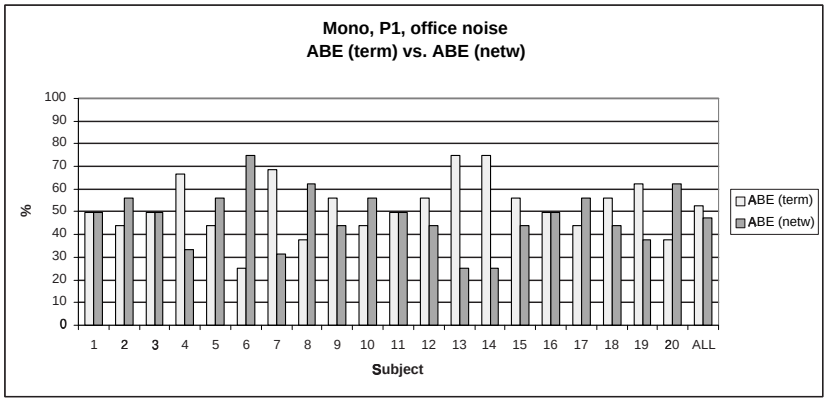

Figure A.11: Preference scores of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), ABE-terminal vs. ABE-network, office noise

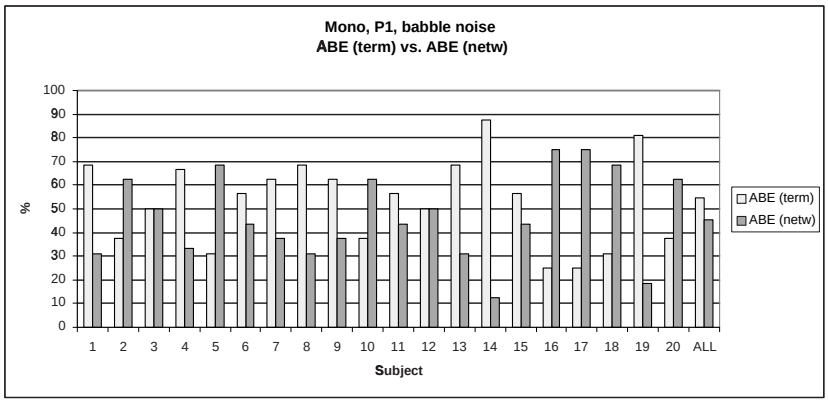

Figure A.12: Preference scores of all the subjects. Mono listening, processing P1 (calls originated from landline telephone), ABE-terminal vs. ABE-network, babble noise

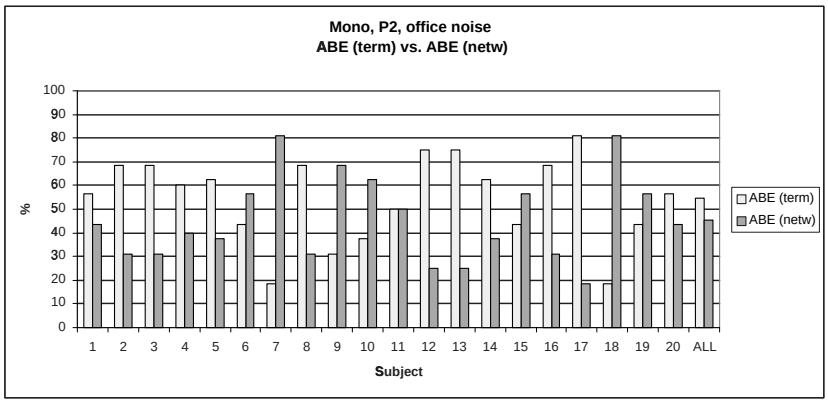

Figure A.13: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), ABE.terminal vs. ABE-network, office noise

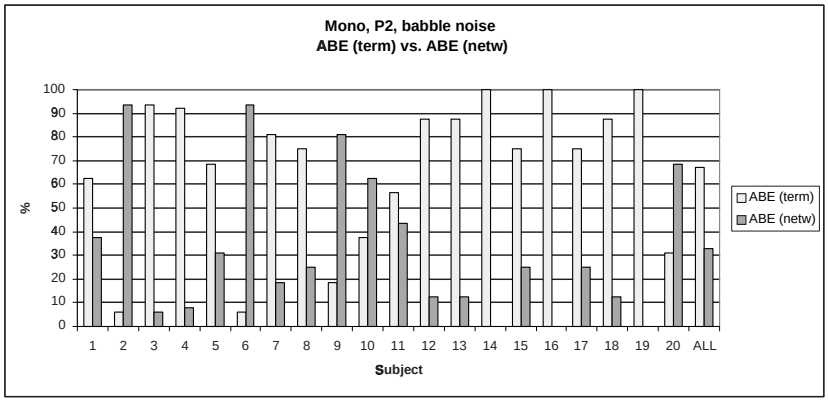

Figure A.14: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), ABE-terminal vs. ABE-network, babble noise

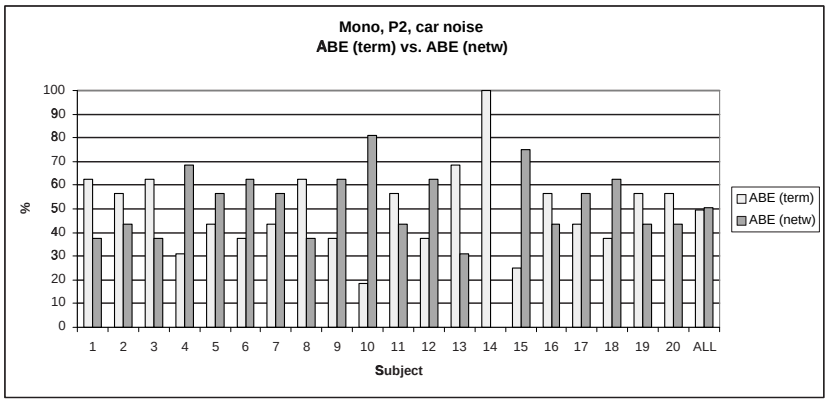

Figure A.15: Preference scores of all the subjects. Mono listening, processing P2 (calls originated from a mobile terminal), ABE-terminal vs. ABE-network, car noise

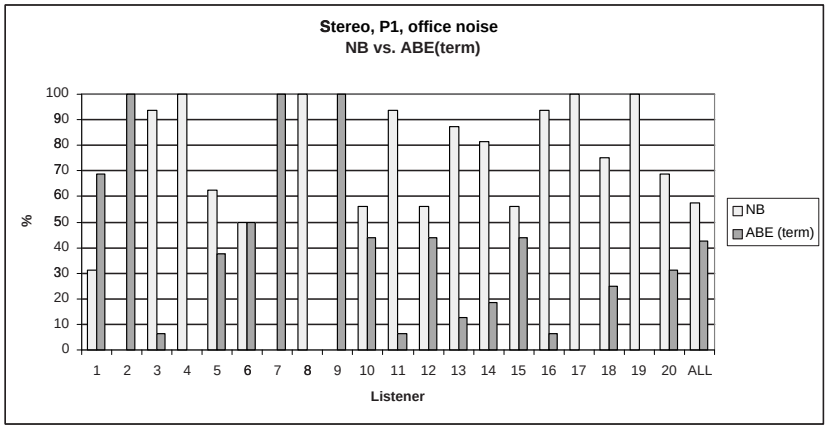

Figure A.16: Preference scores of all the subjects. Stereo listening, processing P1 (calls originated from landline telephone), narrow band vs. ABE-terminal, office noise

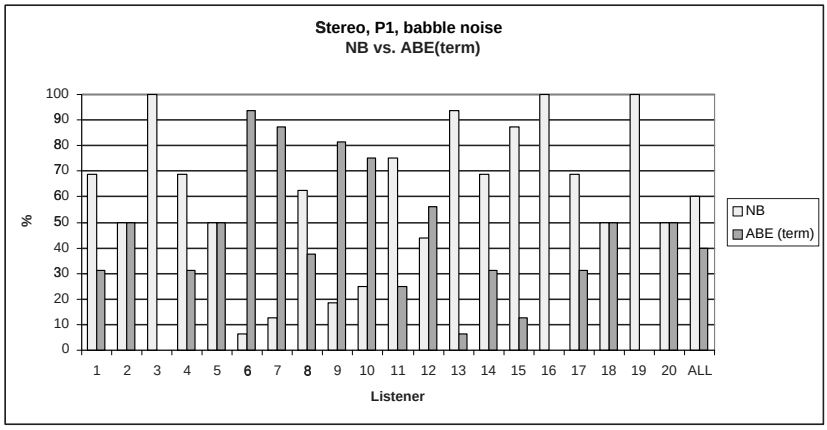

Figure A.17: Preference scores of all the subjects. Stereo listening, processing P1 (calls originated from landline telephone), narrow band vs. ABE-terminal, babble noise

Appendix B

Oral Comments on the Listening Test

Original comments in Finnish

- Liiallinen terävyys on häiritsevää
- Kahdella korvalla kuunneltuna laajakaistaiset huonompia
- Kahdella korvalla kuunneltuna laajakaistaiset parempia
- Valinta tehtävä välillä: hyvä puhe, ikävä kohina - huono puhe, mukava kohina
- Toiset äänistä tutun kuuloisia, toiset "kovia"
- Matalampi taustakohina miellyttävämpi
- Laajoihin voisi tottua
- Osassa äänistä häiritsevää sirinää yläkaistalla
- Kahdella korvalla kuunnellessa kuuli tarkemmin ja silloin tausta ei häirinnyt niin paljoa laajakaistaisissa
- Sirinä ja metalliset äänet ei hyviä
- Laajakaistainen taustamelu luonnollisempi, vaikka "kovempi"
- Puheen selkeys tärkeä kriteeri
- Osa äänistä särähti korvaan
- Kaipaa kahden version puoliväliä

- Selkeät s-äänteet hyviä
- Toiset kuulostivat luonnollisemmilta, mutta valitsi äänen joka muistuttaa puhelimen ääntä

Comments translated to English

- Excessive sharpness is disturbing
- When listening with two ears wideband samples sound worse
- When listening with two ears wideband samples sound better
- The choice has to be made between to cases: good speech, bad noise - bad speech, nice noise
- Some of the sounds sounded familiar, others “hard”
- Lower background noise was more agreeable
- One could get used to wideband sounds
- Some of the sounds included disturbing buzzling in the highband
- With two ears it was easier to listen and then the background noise did not disturb so much in wideband signals
- Buzzling and metallic sounds were not good
- The background noise of wideband signals was more natural even though “harder”
- The clarity of speech is an important criterion
- Some of the sounds were tinny
- Misses the middle version of the two versions
- Clear s-sounds were positive
- Others sounded more natural but chose the sound that reminded of a telephone sound