AALTO UNIVERSITY
School of Science and Technology
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

Emma Jokinen

Adaptive post-filtering of speech in mobile communications

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

Espoo, 03.05.2010

Supervisor: Professor Paavo Alku

AALTO UNIVERSITY
School of Science and Technology
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

ABSTRACT OF THE MASTER'S THESIS

Author:	Emma Jokinen
Name of the thesis:	Adaptive post-filtering of speech in mobile communications
Date:	03.05.2010 **Number of pages:** ix + 61
Faculty:	Electronics, Communications and Automation
Professorship:	S-89
Supervisor:	Prof. Paavo Alku

Speech enhancement is needed to improve the quality and intelligibility of speech degraded by noise. In this thesis, a post-filtering approach for the mobile communication environment was designed. The purpose of this post-processing scheme was to enhance certain frequency regions of speech, so that when it was degraded with a very high level of noise, the speech could still be understood. The post-processing worked by locating the formants of a voiced speech frame by extracting the peaks of the LP spectrum. After this, the first formant was attenuated and the second one enhanced. The idea was to move energy to higher frequencies where the energy level of the noise was lower. The coefficients of the formant filter were optimized with informal listening tests, and the possible tilt of the filter was compensated with a first order low-pass filter. The performance of the post-processing algorithm was studied by analyzing its effects on different voiced sounds and by comparing the filter to other post-filters. It was concluded that the post-processing worked as intended and improved the intelligibility of speech. Some unexpected behavior, such as shifted formants, was also encountered and needs to be further studied. The advantages of this approach are its more adaptive and tunable structure compared to the other methods used for post-processing in high noise levels.

Keywords: Speech enhancement, post-filtering, formant

AALTO YLIOPISTO
Teknillinen korkeakoulu
Elektroniikan, tietoliikenteen ja automaation tiedekunta
Signaalinkäsittelyn ja akustiikan laitos

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Emma Jokinen

Työn nimi: Puheen adaptiivinen jälkisuodatus mobiililaitteissa

Päivämäärä: 03.05.2010 **Sivuja:** ix + 61

Tiedekunta: Elektroniikan, tietoliikenteen ja automaation

Professuuri: S-89

Työn valvoja: Prof. Paavo Alku

Puheen ehostusta tarvitaan kohinaisen puheen laadun ja ymmärrettävyyden parantamisessa. Tässä työssä suunniteltiin matkapuhelimiin tarkoitettu jälkisuodatusalgoritmi. Tämän jälkiprosessoinnin tarkoituksena oli korostaa joitakin taajuusalueita puheessa siten, että sen ymmärtäminen olisi edelleen mahdollista hyvin kovassa kohinassa. Jälkiprosessoinnin alussa soinnillisen puhekehyksen formanttitaajuudet haettiin tarkastelemalla sen LP-spektrissä olevia piikkejä. Tämän jälkeen ensimmäistä löydettyä formanttia vaimennettiin ja toista vahvistettiin. Ideana oli siirtää energiaa korkeammille taajuuksille, jossa kohinan energiataso olisi matalampi. Formanttisuotimen kertoimet optimoitiin kuuntelukokeen avulla ja sen mahdollinen kallistus kompensoitiin ensimmäisen asteen alipäästösuotimella. Lopullisen jälkisuotimen suorituskykyä tarkasteltiin sekä tutkimalla sen vaikutusta erilaisiin soinnillisiin ääntеisiin että vertailemalla suodinta muihin jälkisuotimiin. Saatujen tulosten perusteella voitiin päätellä, että toteutettu menetelmä toimi halutulla tavalla ja onnistui parantamaan puheen ymmärrettävyyttä. Tarkasteluissa tuli kuitenkin ilmi myös yllättäviä piirteitä, kuten formanttien siirtymisiä, jotka vaativat lisätutkimusta. Verrattuna muihin jälkisuodatussysteemeihin, jotka on suunniteltu toimimaan kovassa kohinassa, työssä kehitetyn algoritmin etuna ovat sen adaptiivisuus ja säädettävyys.

Avainsanat: Puheen ehostus, jälkisuodatus, formantti

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Paavo Alku, for his guidance and support during this project. Hannu Pulakka has also been an invaluable help especially in designing the subjective tests, as well as in suggesting new approaches, when I was fresh out of ideas. Furthermore, I would like to extend my gratitude to all other staff at the faculty who have assisted me in various ways during this project.

I would also like to thank all of the kind people, who willingly took part in the listening test. The finalization of this thesis would not have been possible, if it was not for you.

This work has been partly funded by the Nokia corporation, and I am very grateful for this opportunity.

Otaniemi, 03.05.2010

Emma Jokinen

Contents

Abbreviations

3GPP	3rd Generation Partnership Project
ACR	Absolute Category Rating
AGC	Automatic Gain Control
AMR	Adaptive Multi-Rate
CCR	Comparison Category Rating
CMOS	Comparison Mean Opinion Score
DRT	Diagnostic Rhyme Test
ERVU	Energy Redistribution Voiced/Unvoiced
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GI	Gradient Index
GMM	Gaussian Mixture Model
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPAS	Linear Prediction Analysis-by-Synthesis
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MUSHRA	MUlti Stimulus with Hidden Reference and Anchor
PESQ	Perceptual Evaluation of Speech Quality
PDF	Probability Distribution Function
SII	Speech Intelligibility Index
SNR	Signal-to-Noise ratio
SPL	Sound Pressure Level
STL	Software Tool Library
SUS	Semantically Unpredictable Sentences
VCV	Vowel-Consonant-Vowel

List of Figures

List of Tables

Chapter 1

Introduction

Speech is the most important form of human communication. As the field of mobile communications is continuously expanding and evolving, the research of speech becomes increasingly important. The invention of the *linear prediction analysis-by-synthesis* (LPAS) codecs brought a possibility of low rate speech with good or excellent quality. Quality has been even further improved by the implementation of new, wideband codecs which can be used to compress a wideband speech signal. This means that instead of using the conventional telephone band from 300 Hz to 3.4 kHz, the frequency band ranges from 50 Hz to 7000 Hz.

There is constant background noise in the environment, and in mobile communications even the channels and processing generate noise. This means that quite often the speech signal that reaches our ears has already been degraded with interfering noise, regardless of whether the speaker is right next to us or on the other side of the planet. Most of the time the contents of the message can be understood from context, but that is not always the case. Especially in mobile communications where environmental noise often affects both ends of the communication channel, the quality of speech suffers because of noise. However, speech enhancement can be used to minimize the effects of noise.

Speech enhancement covers a wide range of methods which can be used for many different purposes. The most common and obvious application is the area of mobile communications, but it can also be used, for example, in hearing aids as well as helping children in language learning. There is a multitude of different methods which can be roughly divided into two categories based on their basic approach. The first one contains methods that aim at suppressing the noise in the noisy speech signal. The second category consists of schemes that try to enhance some perceptual cues in the speech signal in order to make it stand out more from the noise. Of course, a combination of the two aforementioned approaches is also possible. The goals of different methods can also vary from intelligibility

improvement to quality improvement depending on the setting of the processing and the requirements that the processed speech has to fulfill.

This work focuses on speech enhancement in mobile communications. More specifically, the processing of the speech signal is conducted when it has already reached the receiver's mobile terminal and just before it is played to the listener's ear. This is called post-processing because it happens at the end of the communication chain after the speech has been transmitted over the communication channel and decoded. The opposite would be pre-processing where the speech signal would be modified before encoding at the transmitting device.

The original motivation for this work was to test whether enhancing some of the cues in speech could be used to improve its intelligibility. There are post-processing algorithms in use in the current speech coding standards, but they are ineffective in certain situations, where the noise levels are high. In order to reduce the complexity of the situation to a manageable level, it was assumed that only the receiver, or in this case the listener, would be in a noisy environment. This means that the speech in the receiver's mobile device would be completely noiseless, and the noise would affect the signal only once it would have left the phone. The idea was to modify the speech signal in the mobile device, so that the receiver would be able to hear the processed speech even through the noise.

The thesis is structured as follows: in Chapter 2 the background for the thesis is laid out. First, the concept of linear prediction is explained in brief, and after this, the post-processing of speech and some performance measures are discussed. Chapter 3 consists of a detailed description of the implemented post-processing algorithm. The methods and the results of the subjective tests are presented in Chapter 4, and in Chapter 5 some aspects of the performance are discussed. Chapter 6 concludes the thesis with a discussion of how the post-processing algorithm could be implemented in practice, and, finally, some possible improvements and directions for future research are suggested.

Chapter 2

Background

The purpose of this chapter is to provide the background for the work that was done. First, the concept of linear prediction will be briefly introduced, and then the ways it can be utilized in locating formant frequencies in speech are explained. After this, some previously used approaches to adaptive post-filtering and formant enhancement are presented. Finally, some methods and problems of testing this kind of speech enhancement scenario are discussed, and the setting of this work is defined.

2.1 Linear prediction

Linear prediction (LP) analysis is used in speech processing to form a filter model of the combined effects of the vocal tract, the excitation and lip radiation that can be used to synthesize the sound. This means that the analysis tries to capture the envelope structure of the speech spectrum which, in turn, has been affected by the placement of the articulators, such as the tongue and the lips. The amount of details that are captured depends on the order used in the LP analysis. If it is too high, the resulting LP spectrum resembles the original speech spectrum too much to be of any use. In other words, the effects of the fundamental frequency can still be seen in the spectrum. On the other hand, when the order is too low, the LP spectrum cannot capture all of the necessary details, such as all of the formants. In speech processing the order is generally chosen to be between 10 and 20.

The name linear prediction comes from the fact that we are trying to predict future samples by taking a linear combinition of past samples and some input [1].

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \tag{2.1}$$

In Equation (2.1), s_n is the original speech signal, and coefficients a_k are the linear prediction coefficients which are obtained through the LP analysis. The order of the prediction is

denoted here by p. This is a so called all-pole model which simply means that the resulting synthesis filter structure

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.2}$$

has only poles. To solve the coefficients a_k from (2.1), a least squares approach is used. In this method, the value of the error function is minimized. If the signal is assumed to be deterministic, the error function is

$$E = \sum_n \left(s_n + \sum_{k=1}^{p} a_k s_{n-k} \right)^2 \tag{2.3}$$

By calculating the derivatives of the error function in Equation (2.3) with respect to each of the coefficients a_k, we get p equations which are called the normal equations. When it is specified that the error is minimized over the infinite duration, $-\infty < n < \infty$, they have the form

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \ 1 \leq i \leq p \text{ , where } R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{2.4}$$

This results in the so-called autocorrelation method because the sequence $R(i)$ is the autocorrelation of the signal s_n. The coefficients a_k can then be solved from the Equation (2.4) by direct matrix calculations or by using the *Levinson-Durbin recursion*.

The LP coefficients can be transformed to different forms which can be better suited for different purposes, such as quantization. One of the most often used are the *line spectral pairs* (LSPs) which are also sometimes called *line spectral frequencies* (LSFs). The LSFs are the roots of two polynomials, $P(z)$ and $Q(z)$, which are defined in the following way [2]

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{2.5}$$

$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{2.6}$$

In Equations (2.5) and (2.6), the LP polynomial of order p is denoted by $A_p(z)$. The roots of these polynomials, $P(z)$ and $Q(z)$, always lie on the unit circle in the z-plane, and are interlaced, so that every other root is from $P(z)$ and the rest from $Q(z)$. The advantage of this presentation is that as long as the roots reside on the unit circle and interlace, the corresponding LP synthesis filter is stable. If the LP coefficients were quantized directly, the stability of the filter could not be guaranteed, but when using the LSFs only the ordering of the roots has to be kept constant to ensure stability.

One of the most prominent features of speech are the formants. These are the resonance frequencies of the vocal tract which can be seen as local maxima in the spectrum of voiced speech sounds. The placement of the articulators, such as the lips and the tongue, affects

the formant structure and consequently, the formant frequencies change constantly as the articulators move. The formant frequencies can be used in separating different vowel sounds from each other because they all have specific formant structures.

The results of the LP analysis can be used in locating the formants in two different ways. McCandless calculated the linear prediction spectra from the LP polynomial, and estimated the frequencies of the first three formants from its peaks [3]. The procedure consisted of picking all the local maxima in the spectrum for each speech frame and assigning those to the formants in the most probable way. This means that the continuity of the formant tracks was taken into account as well as the fact that all of the formants reside in certain frequency regions. One major problem with this approach is that sometimes two formants that are close together in frequency produce only one visible peak. On the other hand, all of the local maxima in the linear prediction spectra are not always necessarily formants. Separating the correct formant peaks from all of the other local maxima becomes an increasingly more difficult problem when the order used in the linear prediction grows and more details appear in the spectrum.

Another way of locating formants with the help of the information obtained through LP analysis is by solving the roots of the linear prediction polynomial. In theory, a pair of roots corresponds to a formant peak, but the problem with this method is that not all roots denote a peak in the spectrum. Another difficulty with this approach is that solving the roots of even a 10th order polynomial requires a lot of computation. For this reason, it is mostly used for research purposes where the exact locations of the formants are needed and there are no strict real-time requirements.

A quite common approach is a combination of these two methods. First, the peaks of the spectrum are sought. If there are too many or too few of them or the corresponding frequencies are in frequency regions where formants are not usually found, the roots of the LP polynomial are solved. The poles can then be used to determine which of the peaks are actual formants, or if one of the found formant peaks is actually composed of two formants.

2.2 Post-processing of speech

When the topic of discussion is the post-processing of speech, people are usually thinking of a technical system where the speech is modified in some way after the system has processed it first. However, here post-processing of speech refers to all methods where the speech is modified in order to improve it just before being played to the listener's ear. This broader definition includes also all of the methods where a speech sample is taken from storage, hand-annotated and then some of the perceptually important cues are enhanced.

The goal of post-processing is to improve the quality or intelligibility of the speech sig-

nal. This can mean many things. The profile of the degrading noise can be modified or the speech signal itself can be enhanced or changed in some other way. Methods such as filtering and noise cancellation can be used after the speech signal has been degraded by noise. One way of cancelling the noise is spectral subtraction where the noise power spectrum is estimated and then this estimate is utilized in removing the noise. In this thesis the main interest is in methods that can be used to modify a clean speech signal before it has been corrupted by noise. The idea in this approach is to make the speech more resistant to the degrading noise. This can be achieved by enhancing some of the cues which the auditory system of humans exploits to separate different sounds or by boosting frequency regions, where the energy of the signal is low.

Niederjohn et al. experimented with a method that consisted of first high-pass filtering and then clipping the speech signal in order to make it more intelligible [4, 5]. They aimed at enhancing the second formant in relation to the first one, and studied the effects of this transformation on speech intelligibility. Different values for the filter cutoff frequency and slope were first tested, and the optimum values were found to be 1100 Hz cutoff and 12 dB/octave slope. An intelligibility score was determined for their method by asking ten subjects to recognise words, and then calculating the average score as the percentage of correctly recognized words. When speech processed with the developed method was compared to unprocessed speech, the improvement in intelligibility was found large on almost all of the tested *signal-to-noise ratio* (SNR) values, which ranged from -10 dB to 10 dB.

Hazan et al. used hand-annotated material, and enhanced consonantal regions which contained high densities of acoustic cues [6]. These regions often have low amplitudes and are transient by nature. For this reason, they could be made perceptually more prominent by increasing their amplitudes or durations. Hazan et al. had several approaches for different categories of phones, and they tested the effects these methods had on intelligibility. They conducted two intelligibility tests, one with *vowel-consonant-vowel* (VCV) nonsense words with 13 subjects and the other with *semantically unpredicatable sentences* (SUS) and 32 test subjects. In the first test they achieved an intelligibility improvement of around 10 % with their methods, and it was noticed that the greatest improvements were achieved by enhancing the relative amplitudes of the consonantal regions. In the second test the results did not show clear intelligibility improvements for the sentence materials. Hazan et al. hypothesized that one of the reasons was the difficulty of enhancing the consonants in more varied phonetical contexts.

Skowronski et al. imported some of the features of clear and Lombard speech to normal speech in order to improve its intelligibility [7]. Both types of speech differ from normal speech in many aspects. Clear speech is often produced by the speaker when the listener is hearing-impaired, and is characterized among other things by a slower speaking rate and

more pronounced acoustic cues. Lombard speech, on the other hand, is produced when the speaker is subjected to noise in the speaking environment, and some of its characteristics include larger signal amplitude and energy shift from low frequencies towards higher frequencies.

Skowronski et al. tested an algorithm called ERVU (Energy Redistribution Voiced/Unvoiced) which aimed at boosting the consonant-vowel ratio. The algorithm worked by first recognising the unvoiced regions and then amplifying them. After this, the energy of the signal was calibrated to the level where it was before the processing. This procedure effectively moved energy from the voiced regions to the unvoiced regions. The performance of ERVU was tested for intelligibility improvement against simple high-pass filtering with listening tests. The test was a forced two choice test with words from confusable sets and 25 test subjects. A confusable set consists of words that only differ very little and can thus be easily confused with each other. An example of such a group of words is ”too”, ”two” and ”to”. Skowronski et al. concluded from the results that both the high-pass filtering and ERVU improved intelligibility in most cases, and suggested that combining the two approaches could be used in achieving even larger gains in intelligibility.

Hall et al. studied intelligibility enhancement in situations where there is babble noise in the listener's end [8]. They experimented with differentiation and formant equalization, which are both similar to high-pass filtering. In differentiation the processed signal y_n was formed by

$$y_n = x_n - 0.995x_{n-1} \ , \tag{2.7}$$

where x_n is the original speech signal. Formant equalization was done by taking the average formant amplitudes of male speakers and equalizing the speech signal with a filter that had these amplitudes reversed. This effectively attenuated the first formant and enhanced the second formant.

Hall et al. first studied the possible intelligibility improvement of the methods with the *diagnostic rhyme test* (DRT), and concluded that the processing increased intelligibility in both cases. After this they conducted a *comparison category rating* (CCR) test. The *comparison mean opinion scores* (CMOS) for the tested methods versus no processing were positive except with the highest SNR value (6 dB). They concluded that both of the processing approaches were preferred over no processing. Out of the two methods, formant equalization was preferred.

In adaptive post-filtering, we have a filter structure that updates itself based on the changes in the speech signal. The conventional post-filter structure presented by Chen at al. consists

of three separate filters in cascade [9].

$$H(z) = G\,G_l\,\underbrace{\frac{1+\gamma z^{-p}}{1-\lambda z^{-p}}}_{=H_l(z)}\,\underbrace{\frac{1-P(\frac{z}{\beta})}{1-P(\frac{z}{\alpha})}}_{=H_s(z)}\,\underbrace{(1-\mu z^{-1})}_{=H_t(z)} \tag{2.8}$$

In Equation (2.8), $H_l(z)$ is called the long-term post-filter, $H_s(z)$ the short-term post-filter and $H_t(z)$ the tilt filter. The long-term post-filter uses the information about the pitch period p to emphasize the pitch harmonics. It also attenuates the valleys between pitch harmonic peaks, and reduces the noise in these frequency regions. The coefficients γ and λ were determined based on the degree of voicing in the speech segment.

The short-term post-filter is based on the LP coefficients which contain the overall structure of the speech spectrum, as mentioned before. It is used to enhance the formant peaks, and to attenuate the valleys between them. This is achieved by moving the poles and zeros of the short-term post-filter further apart inside the unit circle in the z-plane by changing the values of the constants α and β, so that $0 < \beta < \alpha < 1$. The stability of the filter is always guaranteed as long as the poles stay inside the unit circle. When the poles are moved closer to the unit circle, the peaks in the spectrum become sharper and narrower. If the difference between α and β is large enough, the peaks become more prominent, and are easier to locate from the spectrum.

After the speech was processed with the post-filter, the power of the speech segment had to be adjusted to the level where it was before the processing took place. This was done with an *automatic gain control* (AGC) algorithm that estimated the powers of both the original and processed speech, and then scaled the filtered speech with a suitable value determined from the power estimates.

Finally, the processed speech was filtered with the spectral tilt filter which was a simple high-pass filter. It compensated for the low-pass effect caused by the short-term post-filter, and reduced the audible muffling in the processed speech. The whole filter structure is called adaptive because it changes with the characteristics of the speech segment that is being processed. For example, the pitch and the formant structure of the current frame affect the structure of the post-filter. The constant coefficients used in the filters were determined by informal listening tests, and their optimal values were $\alpha = 0.8$, $\beta = 0.5$ and $\mu = 0.5$.

The short-term post-filter and the tilt filter defined in the narrowband *adaptive multi-rate* (AMR) standard [10] have exactly the same structure as in Equation (2.8), only the constants are different. The values of α and β depend on the quality of the channel, so that when the SNR decreases, the effect of the post-filtering increases. This means that the post-filtering becomes stronger as the channel deteriorates. The parameter of the first degree tilt filter, μ, is made adaptive and depends on the impulse response of the short-term post-filter. The standard does not use a long-term post-filter, and it should also be noted that

the LP coefficients that are required in the post-filter are not calculated at the receiver but transmitted as a quantified version from the transmitting side. Chen et al. do not mention how these values are obtained in order to keep the discussion as general as possible [9]. The AGC algorithm used in the AMR standard processes each frame of 20 ms in four subframes with a length of 5 ms, and the calculation of the scaling factor itself is very similar to the algorithm used with the conventional post-filter. A more detailed description of this algorithm is given in Chapter 3.

Mustapha et al. estimated the locations and the bandwidths of the formants based on the roots of the LP polynomial [11]. This information was then used to construct a modified Yule-Walker filter which was fitted with a least squares approach. Their post-filter had a flat overall frequency response meaning that no separate tilt filter was needed. It was used in attenuating the valleys between formants while maintaining the peaks unaltered. The performance of the post-filter compared to the conventional post-filter in [9] was studied with two types of subjective listening tests. The first test was an *absolute category rating* (ACR) test with 24 listeners and it gave a higher *mean opinion score* (MOS) value to the new post-filter. The second test that was conducted was a CCR test with 16 listeners. The new post-filter was again compared to the conventional one with 4 kb/s speech coders. From the results, Mustapha et al. concluded that on average the subjects preferred the new post-filter.

Kim et al. used a short-term post-filter that was also designed to compensate for the spectral tilt [12]. Therefore, a separate tilt filter was no longer needed. The coefficients obtained from LP analysis were used to derive the coefficients of the post-filter along with the coefficients from the LSP polynomials. Their short-term post-filter had the following form

$$H_s(z) = \frac{P_M(\frac{z}{\alpha_1})Q_M(\frac{z}{\alpha_2})}{A_M(\frac{z}{2\beta})} \tag{2.9}$$

In Equation (2.9), the polynomials $P_M(z)$ and $Q_M(z)$ represent the LSP polynomials derived from the Mth order LP polynomial, $A_M(z)$. The values of α_1,α_2 and β were determined based on subjective listening tests and chosen optimally, so that $\alpha_1 + \alpha_2 = 2\beta$ to minimize the phase distortion.

Kim et al. studied the perfomance of their filtering scheme with subjective listening tests. The test was an CCR test against the conventional post-filter with eight listeners. They concluded that their scheme was able to reach the same quality as the reference post-filter. However, the advantage was that the order of the their post-filter could be lowered, and no additional tilt filter was needed. These differences might be useful in lowering the computational complexity of the post-processing algorithm.

Farsi designed a post-filter that was based on factorizing the LP synthesis filter to separate

cascaded filters, so that the main formant could be processed separately [13]. The idea was to make the main formant bandwidth narrower and to maintain other formant information while attenuating the valley regions between formants. This was achieved by first calculating the poles of the synthesis filter and then dividing the poles amongst the formants. This means that the poles that were affecting the shape of the formant, were assigned to that formant. Then the poles associated with the main formant were manipulated either by changing their radii closer to the unit circle or by changing the angles of the poles closer to one another. The optimal change was determined by considering the effects it had both on the other formants and on the valleys between them.

The performance of the post-filter was tested with subjective listening tests. It was compared in a CCR test against the post-filter in the ACELP standard [14] which is similar to the conventional post-filter. The CMOS value showed that the 14 test subjects preferred the developed post-filter.

Grancharov et al. aimed at improving the performance of the short-term post-filter by making it adaptive to the statistics of the degrading noise [15]. Their basic idea was to track the *spectral flatness* of the noisy speech signal and to determine the post-filter coefficients from these values. The correspondence between spectral flatness and the filter parameters was derived by an offline training algorithm.

For the training they created noisy speech samples with different types of noise and varying SNR. The optimal post-filtering parameters for each noise condition were found by processing each sample with multiple different parameter combinations and choosing the one that had the minimum distance from the clean speech sample. The distance measure was based on a perceptual model presented in [16] by Dau et al. A nonlinear correspondence was derived between the spectral flatness values and the optimal parameters.

Grancharov et al. compared the performance of the constructed post-filter against the conventional post-filter in [9] with and without pre-processing noise suppression. They conducted subjective listening tests and some objective measurements with different noise types and varying SNR. In the latter, *perceptual evaluation of speech quality* (PESQ) was used to approximate the subjective quality of the speech processed with the post-filters. Another objective measure, based on Dau's perceptual model, was used to measure the difference between processed speech with different post-filters and the original speech signal. Based on the objective measures both the designed post-filter and the conventional post-filter with noise suppression outperformed the conventional post-filter. However, they were not able to determine a clear winner. The subjective test was conducted with a MUSHRA (multi stimulus with hidden reference and anchor) type setting. In this test eight listeners were asked to rate the signals on a scale of 0-100. Grancharov et al. concluded from the average scores that the proposed post-filter was better than either of the reference systems.

Chen et al. proposed a new perceptual post-filter in [17] and [18]. Their idea was to estimate the gains of the perceptual post-filter with a *minimum mean square error* (MMSE) estimator. They used *Gaussian mixture models* (GMMs) to model the *probability density function* (PDF) of the feature vectors which consisted of information obtained at the encoder about the speech frame. The pdf was needed in the MMSE estimation. The parameters for the GMMs were obtained with training in the encoder, and the actual post-filter was estimated in the decoder. The performance of the perceptual post-filter was evaluated with informal listening tests. The speech was coded with the G.723.1 codec and the new post-filter was compared against the one defined in the standard [19]. They did not, however, present any results from these tests, but merely concluded that the proposed post-filter gave more natural speech, while maintaining intelligibility.

2.3 Performance measures

Some methods for measuring the performance of a post-processing system for speech have been mentioned earlier in this chapter. Different speech signals can be compared, for example, in terms of quality, distortion or intelligibility. It should be noted that these measures are related in many ways. Distortion means here the differences in the time or frequency domain signal which are caused by the processing under evaluation. A distortion measure can be determined in an objective way simply by calculating the distance between the original and the processed signal. The definition of the distance measure itself can range from simple *Euclidean distance* to the Dau measure mentioned earlier. The problem with these distortion measures is the fact that they do not measure audible distortions very well. Even though two time domain speech signals have a large distance, the difference can go undetected by human listeners. On the other hand, a small peak at the right place can be extremely annoying. These effects are due to the complexity of the hearing mechanism. Some fairly simple objective measures, such as *log-likelihood ratio* (LLR), *frequency-weighted segmental SNR* and cepstrum distance seem to give some indication of the perceived quality, but only with a correlation of 0.64 at the highest [20].

A more complex objective measure is the PESQ score [21] which predicts the subjective quality of a speech signal. It works by first aligning the degraded and the original signals in the time domain and then calculating their difference with a perceptual model. After this, the distortion measures are transformed to a MOS score with a cognitive model. In [21], the correlation between the MOS scores obtained from PESQ and the ones obtained from subjective tests was 0.935. However, in [20] it was found to be only 0.65. The need for subjective quality tests is still evident.

There are many types of subjective listening tests that can be used in measuring speech

quality [22]. In ACR tests, the subjects are asked to rate the samples they have heard on a 5-point scale such as in Table 2.1(a). The experiment should include at least one reference signal, so that tests conducted in different locations can be compared to each other. The average score obtained from combining all the listeners is called the MOS score.

Another type of test is the CCR test which can be used to compare two processing methods or speech codecs. In the test, two samples are played to the listener succesively, and the listener is asked to rate the quality of the second sample compared to the first. The scale is a 7-point scale with adjectives and correponding scores as in Table 2.1(b). To remove the effects of the ordering of the samples from the results, the pairs are ordered randomly, so that half of them have the reference sample first, and the other half starts with the processed sample. The total average score obtained from the test is called a CMOS score, and it is only meaningful when indicated what was used as reference.

A MUSHRA [23] test setting has been primarily intended for measuring the quality of low quality broadcasting systems, but the same ideas can also be used for evaluating speech. In the MUSHRA setting, the listener can compare multiple signals to each other and to a reference signal while adjusting the ratings for the signals. There should be also one hidden reference and one anchor which can be used to check that the listener is consistent and to compare the results to those obtained from other similar tests. The rating is similar to the ACR test, but the scale is continuous and ranges from 0 to 100.

Several approaches exist also for estimating the intelligibility of a speech signal, both subjective and objective. In a diagnostic rhyme test [24], the subjects are asked to pick the word that they heard from two rhyming options which are presented in writing. The percentage of words recognised correctly can be used as an intelligibility score for the method under testing.

Table 2.1 – Rating scales given in ITU-T Rec. P.800 [22].

(a) Absolute category rating.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

(b) Comparison category rating.

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

The SUS [25] test was mostly designed for measuring the intelligibility of text-to-speech systems, but it has also been used in the context of post-processing of speech. The sentences are constructed from a list of words with the help of certain basic grammatical rules. The idea is that the sentences are such that individual words cannot be determined from their context. The intelligibility score can be calculated as the percentage of the words that the test subjects recognized correctly.

The *speech intelligibility index* (SII) is an objective method for measuring the intelligibility of speech [26, 27]. It works by measuring the speech spectrum level, the noise spectrum level and the hearing threshold level, and then determining the *band audibility function* based on the results. This function tells the percentage of the speech range which can be heard on the different frequency bands. The final speech intelligibility index is obtained by weighting the band audibility function with a *band importance function*, and then summing over the whole frequency range. The result should be a value between zero and one where zero means completely unintelligible and one is perfect intelligibility.

The problem with post-processing of speech is that the actual information signal itself is often modified to produce better intelligibility, and it can also have an effect on the quality of speech. The processed signal can start sounding unnatural and irritating. As a result, the methods used for post-processing of speech with the goal of improving its intelligibility should be tested for both intelligibility and quality. It is hard to combine these two types of test into one, and often just a MOS score or an intelligibility score is determined with subjective tests. Another approach is conducting two separate tests, one for intelligibility and one for quality. Both of these methods have been used in the articles presented before in this chapter.

2.4 Focus of this work

In this thesis the concentration is on refining a short-term post-filter which adapts itself to changing conditions and is designed to operate in a mobile phone. This means that the algorithm should be automated and operate with very little delay. Therefore, the necessary computations have to be quite simple and straightforward, and the speech material cannot be hand-annotated before processing. In other words, the algorithm has to be able to determine independently when and what kind of processing should be done.

The speech is processed after it has been encoded and decoded by the AMR narrowband coder, but before the noise has corrupted the signal. This corresponds to the real life situation where the listener listens to clean speech through a mobile phone in an extremely noisy environment. The method was designed to improve the intelligibility of the speech signal without causing too much distortion or audible artifacts.

Chapter 3

Implementation

The focus of this chapter is the approach to adaptive post-filtering that was formulated in this thesis. First, the general setting of the problem will be discussed, and then the actual implementation of the algorithm itself will be presented in greater detail. Finally, the implementation for the calculation of the speech intelligibility index is introduced. All of the work was done using the MATLAB software.

3.1 The general setting

The problem setting is similar to the one defined by Sauert et al. in their article on near end listening enhancement [28]. They also have a situation where the receiver of the speech signal is affected by serious degrading noise, and the output power of the speaker of the phone is limited. This means that the listener cannot simply increase the volume in order to overcome the effects of the environmental noise. The setting used in this work is depicted in Figure 3.1.

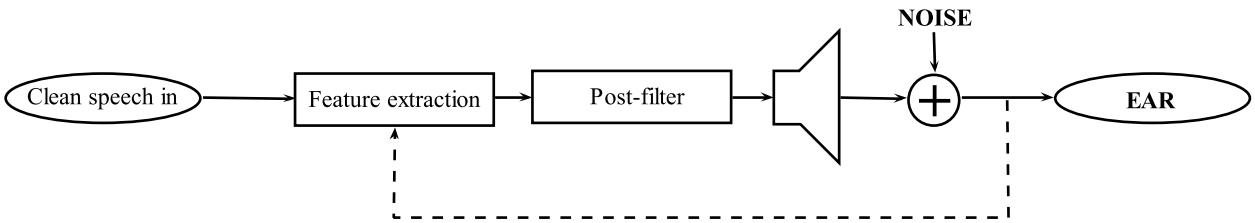


Figure 3.1 – The setting of the problem.

In the figure, there is clean speech coming in to the feature extraction block. In reality, the speech signal has been quantized, encoded and sent over the communication channel before decoding it at the receiver's mobile terminal. It is not the original speech signal, but

it is referred to as clean speech because it is assumed that the speech is not degraded by noise at the transmitting side and that the channel does not distort the signal.

The feature extraction and the post-filter blocks form the post-processing part. These are done just before the speech is played out through the loudspeaker. When the signal reaches open air, it is corrupted by environmental noise, and after this, it is heard by the receiver. In the figure, there is also a feedback loop marked with dashes. This can be used, for instance, to adapt the post-processing to the characteristics of the noise.

Before post-processing

As mentioned earlier, in this scenario, the speech signal that reaches the post-processing block differs from the original that was uttered at the transmitting side. In order to simulate this, the speech samples have to go through some processing. The whole sequence is presented in Figure 3.2.

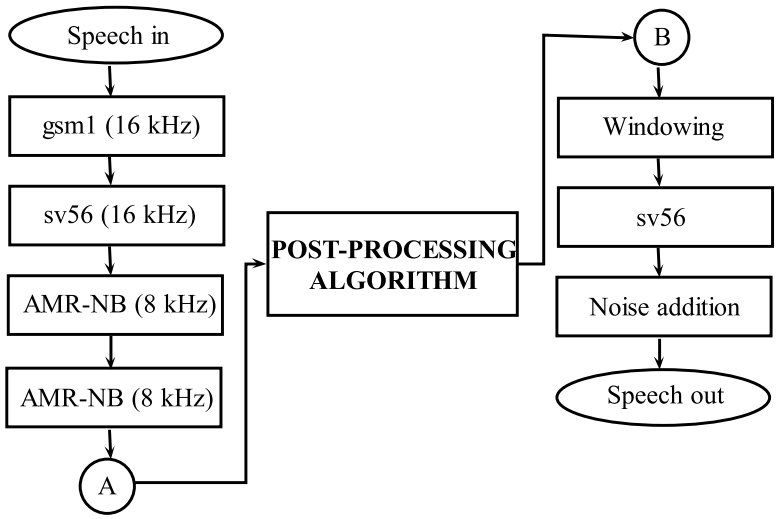


Figure 3.2 – Overview of the processing.

First, a clean speech sample is read into MATLAB and resampled, so that its sampling frequency is 16 kHz. After this, it is filtered with the `gsm1` filter which can be found from the *software tool library* (STL) provided by ITU-T [29, 30]. The function of this filter is to mimic the input filtering of a mobile phone. The lower frequencies from 50 to 300 Hz of the speech signal are suppressed using a high-pass filter with a cut-off frequency around 200 Hz.

After the filtering, the active power level of the speech sample is set to -26 dBov with the method defined in ITU-T standard P.56 [31]. The power level is defined relative to a reference power level of 1.0 and the ”ov” refers to *digital overload signal level* [32]. The

idea of the algorithm is to seek out the silence intervals from the speech signal and to leave those out when calculating the speech level. A function for this purpose, called `sv56`, can also be found from the ITU-T STL.

Before the level equalized speech reaches the actual post-processing block, it is twice encoded and decoded with the AMR narrowband codec [10]. This simulates the use of a mobile communication channel where the information is sent over the channel in AMR coded form. The encoding-decoding operation is done twice to take into account the worst case scenario of tandem coding where the data is sent over the public switched telephone network at some point. This can potentially degrade speech quality even further. The AMR encoder and decoder functions were obtained from the *3rd Generation Partnership Project* (3GPP) ANSI C realization [33], and they were used at the highest bit rate of 12.2 kbit/s.

After post-processing

When the speech signal has passed through the post-processing block, it is at node B in Figure 3.2. First, the beginning and the end of the processed signal is smoothed with half of a 10 ms long Blackman-Harris window which can be seen in Figure 3.3. If there are some peaks in the signal during these silent periods, the windowing will suppress them. In reality, windowing of this kind, which removes breathing noises and other natural sounds, would not be used. However, here the speech is processed in samples which are parts of a larger entity instead of having a continuous flow of data.

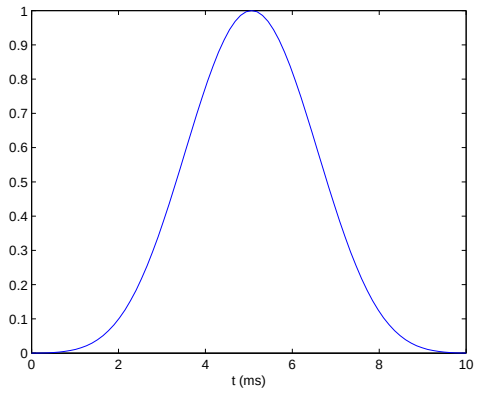


Figure 3.3 – The Blackman-Harris window used for smoothing the beginning and the end of the processed samples.

After this, the speech is once again level-equalized to -26 dBov with the `sv56` algorithm. Finally, the noise is added to the speech signal. The noise type in use is car noise which is very stationary. In other words, the noise profile does not vary much over time. Figure 3.4

presents the spectrum of the noise, and it shows that most of the energy is concentrated on low frequencies. Before adding the two signals together, the noise signal is scaled, so that the resulting noisy speech signal will have the desired SNR value.

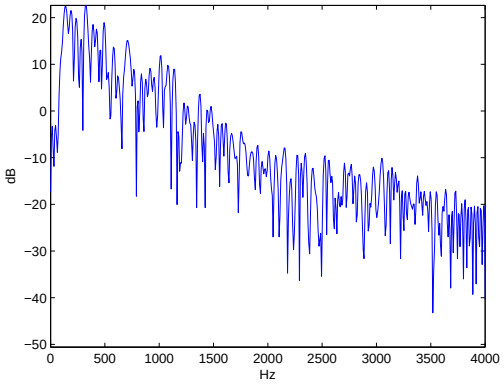


Figure 3.4 – Car noise magnitude spectrum.

3.2 The post-processing algorithm

In the post-processing block, the speech is processed in frames of 20 ms with a sampling frequency of 8 kHz. The frames are extracted from the signal using a rectangular window. The flowchart of the algorithm that is used to process the frames is presented in Figure 3.5. As can be seen from the figure, also information about the next frame is needed during the processing, and, therefore, 20 ms of additional delay is introduced.

Before the different blocks in the flowchart are introduced more closely, the basic logic of the algorithm is presented. At first, the current and the next frame are processed through the first three blocks in the flowchart. After this, there are four different courses of action to take depending on whether the frames are classified as voiced or not. If a frame is not voiced, it means that it is either silent or unvoiced. Here the term *unvoiced* will be used to refer to both cases. The post-filtering is only done on voiced speech frames and the filter coefficients are interpolated between consecutive voiced frames. The transition between unvoiced and voiced frames is smoothed.

If both the current and the next frame are voiced, the algorithm proceeds normally, and the coefficients of the filters are interpolated linearly between the frames. If neither is voiced, then nothing is done, and the processing is moved to the next frame. If only the current frame is voiced, there is no need for interpolation, and the whole frame is filtered using the same coefficients. If the situation is the exact opposite, meaning that the current frame is classified as unvoiced while the next frame is voiced, a smoothing period is used.

This means that part of the current frame is filtered in order to avoid sudden transitions between the frames.

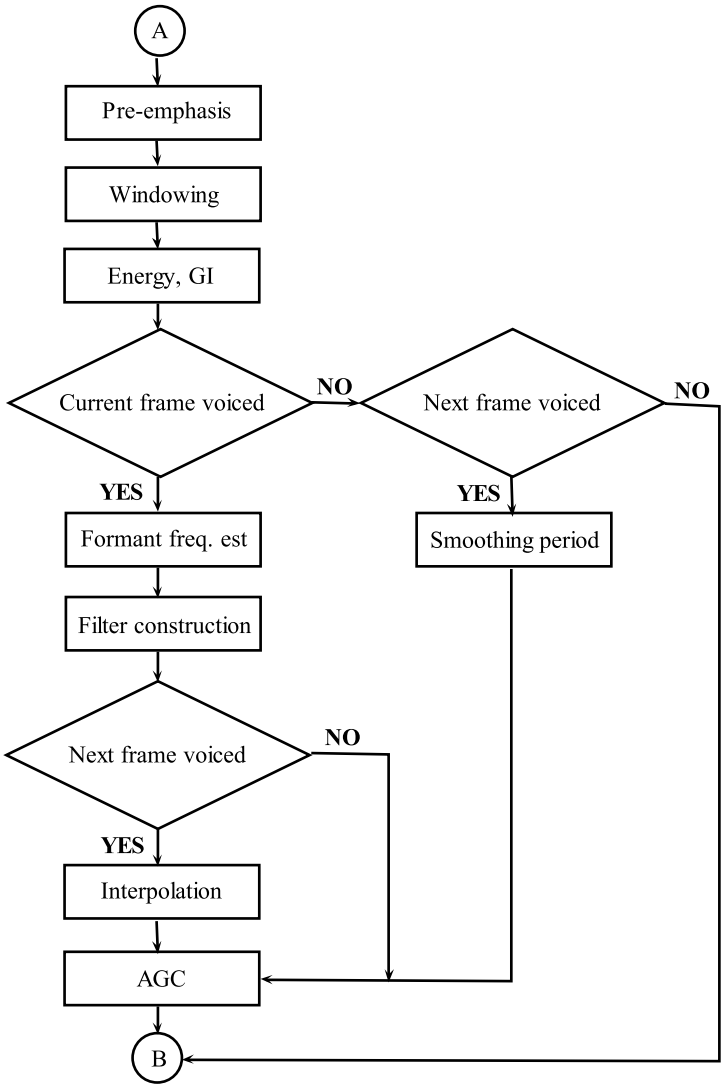


Figure 3.5 – The flowchart of the post-processing algorithm.

Pre-emphasis and windowing

Before anything else, the speech frame is pre-emphasized and windowed for the purposes of the LP analysis. Pre-emphasis is used because it helps in bringing out more details from the higher frequency bands more easily in the linear prediction analysis [34]. Windowing, on the other hand, smoothes the transitions between consecutive frames.

The pre-emphasis is done using a filter which is defined in the following manner.

$$H(z) = 1 + a_1 z^{-1} \,, \tag{3.1}$$

where a_1 is the first order LP coefficient. For the windowing, a regular Hamming window is used.

Voiced/Unvoiced decision

Energy of the frame is used in separating silent periods from active speech. Some of the unvoiced sounds, namely sibilants, have such low energies that they can also be separated from voiced sounds with this measure. The energy of the frame alone is not enough to separate between all unvoiced and voiced phones, and therefore other simple features were looked into. In articles [35] and [36], the authors utilize multiple features, such as the *gradient index* and the *frame energy ratio*, and a few of these were implemented and tested. Stop consonants, such as [p], [t] and [k], turned out to be especially difficult to separate with a simple approach. In the end, the gradient index was chosen as the other measure. For voiced sounds, it tends to have a low value whereas unvoiced sounds have higher values [37]. It can be calculated for frame n as in (3.2).

$$x_{gi}(n) = \frac{\sum_{\kappa=1}^{N_\kappa - 1} \Psi(\kappa) \left| s_\kappa - s_{\kappa-1} \right|}{\sqrt{\sum_{\kappa=0}^{N_\kappa - 1} (s_\kappa)^2)}}, \tag{3.2}$$

where N_κ is the frame size, s_n is the speech signal and $\Psi(\kappa)$ is defined as

$$\Psi(\kappa) = \frac{1}{2} \left| \psi(\kappa) - \psi(\kappa - 1)) \right| , \; \psi(\kappa) = \begin{cases} -1 & \text{, when } s_\kappa - s_{\kappa-1} < 0 \\ 0 & \text{, when } s_\kappa - s_{\kappa-1} = 0 \\ 1 & \text{, when } s_\kappa - s_{\kappa-1} > 0 \end{cases} \tag{3.3}$$

Thus a frame is classified as voiced, if its gradient index value is lower than GI_{limit}, and its energy is above the limit, E_{limit}. The constants were determined by testing with speech material, and the values that were chosen for the decision were $GI_{limit} = 8$ and $E_{limit} = 2e - 4$.

Formant frequency estimation

The most important requirement for the approach to formant frequency estimation was speed. Because the post-processing algorithm is meant to work in a mobile phone, additional delays should be avoided. It was also noticed that the exact locations of the formants were not absolutely necessary because most errors did not affect the audible quality of the speech. For these reasons, the peak picking approach was chosen.

First, the LP coefficients of the frame are calculated using tenth order linear prediction. Normally, the formant frequencies are estimated by picking the peaks of the LP spectrum. Because these peaks are sometimes hard to detect, a conventional post-filter structure based on the LP coefficients is studied instead. It can be seen as an enhanced LP where the formant peaks are sharper and more pronounced in the spectrum.

$$H_{enh}(z) = \frac{1 - P\left(\frac{z}{0.9}\right)}{1 - P\left(\frac{z}{0.99}\right)} \tag{3.4}$$

In Equation (3.4), $P(z)$ is the LP polynomial. The amplitude response of the post-filter, $H_{enh}(z)$, is calculated with a 256-sample FFT. The first three peaks are determined, and their frequencies are compared to the formant locations in the previous frame. The peaks that are closest to the formants of the previous frame are selected. If none of the found peaks are close enough, the frequency values from the previous frame are used instead. The estimated frequencies of the formants are allowed to change at most 50 Hz between consecutive frames. This value was obtained through experimentation with speech samples.

Post-filter structure

The structure of the filter that is used to modify individual formants is presented in Equation (3.5).

$$H(z) = \frac{1 - 2r_1 cos(\theta) z^{-1} + r_1^2 z^{-2}}{1 - 2r_2 cos(\theta) z^{-1} + r_2^2 z^{-2}} \tag{3.5}$$

The constant θ is the angle of the formant in radians, and the values of r_1 and r_2 determine whether the filter is enhancing or attenuating the formant and by how much. If $r_1 < r_2$, then the effect of the post-filter is in fact one of enhancement, but if the opposite is true, the post-filter will have a dampening effect. If the constants are far apart, the post-filter will be very wide and two values close to each other will produce a sharper form. Figure 3.6 shows three different filters with the same value of θ, but different r_1 and r_2.

In the post-filter structure, the first two formants are manipulated. The complete formant filter is presented in Equation (3.6).

$$H_{pf}(z) = \frac{1 - 2 \cdot 0.9 cos(\theta_1) z^{-1} + 0.9^2 z^{-2}}{1 - 2r_1 cos(\theta_1) z^{-1} + r_1^2 z^{-2}} \cdot \frac{1 - 2 \cdot 0.9 cos(\theta_2) z^{-1} + 0.9^2 z^{-2}}{1 - 2r_2 cos(\theta_2) z^{-1} + r_2^2 z^{-2}} \tag{3.6}$$

In order to reduce the amount of possible parameters, the numerator parameters in both filters were attached to the value 0.9. The two remaining constants are chosen, so that the first formant is attenuated, $r_1 \leq 0.9$, and the second formant is enhanced, $r_2 \geq 0.9$. This moves some of the energy from the lower frequency bands to the higher frequencies. When most of the noise energy is in the lower frequencies, this operation enhances vocal cues thus making the speech easier to understand. The constant parameters r_1 and r_2 were

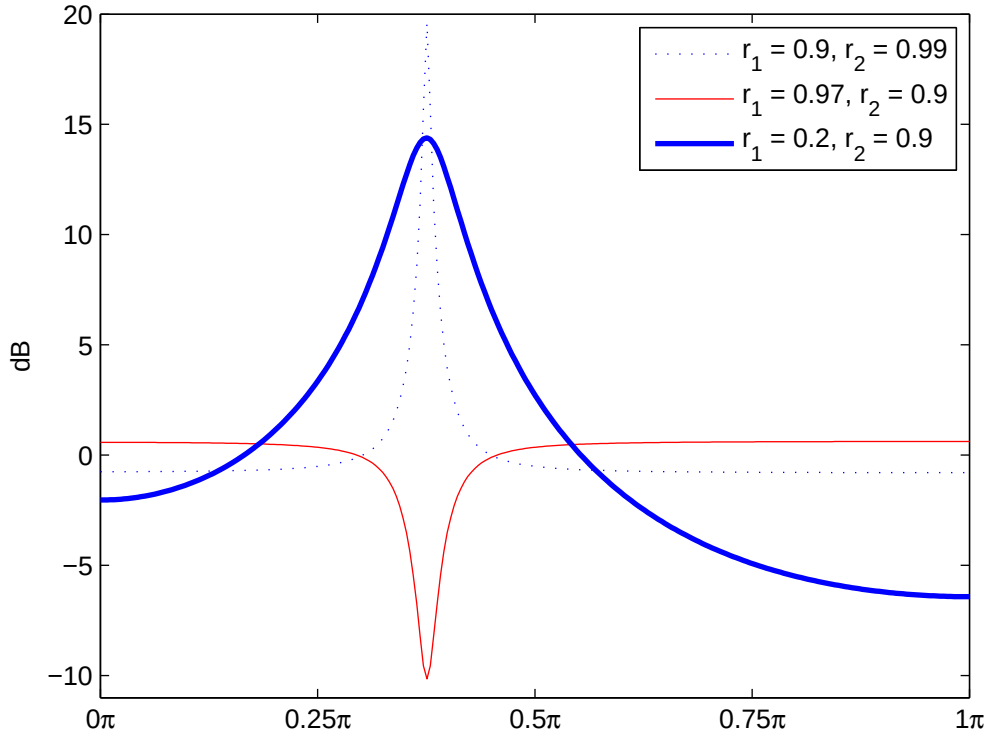


Figure 3.6 – Effects of the parameters on the filter amplitude response.

determined by subjective listening tests which are described in detail in the next chapter. Their final values are $r_1 = 0.46$ and $r_2 = 0.93$.

Tilt compensation

If the first formant is strongly attenuated, the post-filter has an overall structure that resembles a high-pass filter. On one hand, this is a desired effect, but if the higher frequencies have a strong tilt, the quality and the naturalness of the speech signal are reduced. In order to completely avoid or at least diminish this effect, a tilt compensation filter was added to the structure.

$$H_{tilt}(z) = \frac{1}{1 - \mu z^{-1}} \tag{3.7}$$

In Equation (3.7), the parameter μ is the coefficient of the first order linear prediction of the post-filter $H_{pf}(z)$. The parameter contains some information on the overall tilt of the post-filter. The compensation filter can thus adapt to changes in the parameters r_1, r_2 and θ_i. A first order filter cannot completely compensate for the tilt, and for this reason the

range of the attenuation of the first formant was limited, so that $r_1 \geq 0.27$. The amplitude response of the whole post-filter structure with $H_{pf}(z)$ and $H_{tilt}(z)$ in cascade is derived in Appendix A.

The effects of the compensation filter can be seen from Figure 3.7. It presents the amplitude response of the post-filter with and without compensation. Both cases have the same parameter values, $r_1 = 0.3$ and $r_2 = 0.95$, and the formant frequencies are selected to be 450 Hz and 2200 Hz. The compensation filter reduces the attenuation of the first formant, but, most importantly, it lowers the enhancement after the second formant.

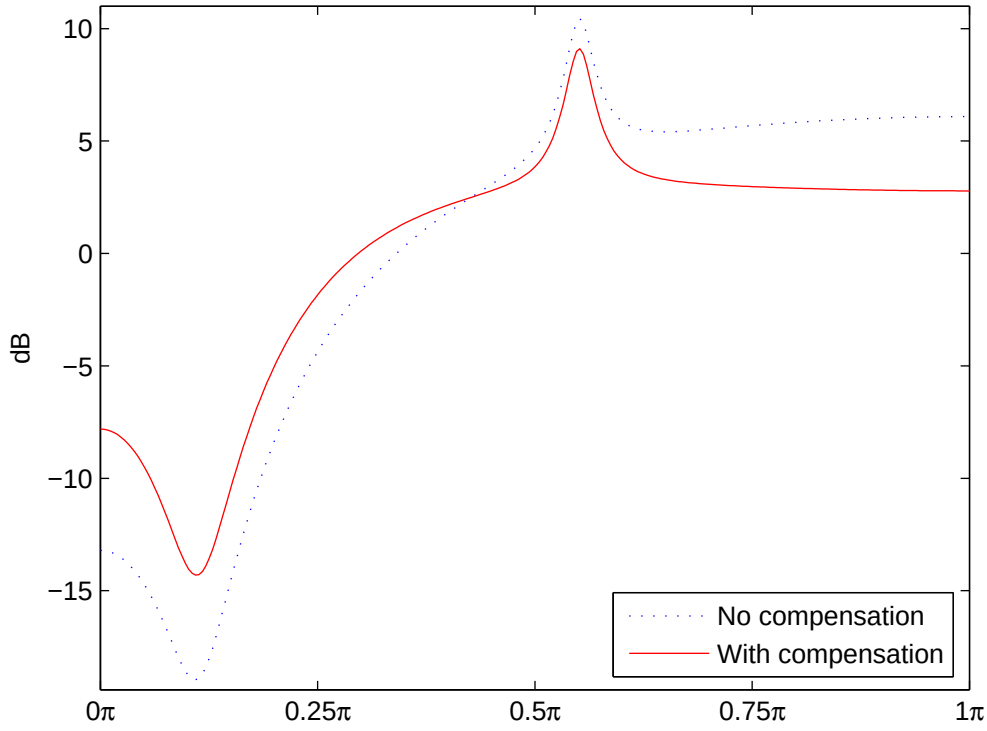


Figure 3.7 – The effects of the compensation filter.

Interpolation

In order to avoid artifacts caused by sudden transitions between consecutive voiced frames, the coefficients of the post-filter are interpolated between the current frame and the next frame. At first, the interpolation was done on a sample-by-sample basis, so that the best possible quality would be achieved. It was, however, noticed that there was no audible

difference when the interpolation was done only every 20th sample. This change also significantly reduced the processing delay.

The coefficients of the filters are first transformed to the LSF domain and then interpolated linearly. The transformation to line spectral frequencies is done to ensure that the filter remains stable even though its coefficients change. As was discussed earlier in Section 2.1, the stability of the filter is guaranteed much more easily in the LSF domain. The filter coefficients for a subframe of 20 samples are obtained in the following manner :

$$a_{sf}(n) = \left(1 - \frac{i}{N-1}\right) \cdot a_{cf}(n) + \frac{i}{N-1} \cdot a_{nf}(n) \tag{3.8}$$

In Equation (3.8), a_{sf} denotes the subframe coefficients, a_{cf} the coefficients of the current frame and a_{nf} those of the next frame. The length of the frames is N, and the starting index of the subframe inside the larger frame is i, $0 \leq i \leq N-1$. Both the numerator and denominator coefficients for the subframe filter are interpolated separately.

Smoothing period

When the current frame is classified as unvoiced and the next frame is voiced, a smoothing period of 5 ms is used to avoid sudden transitions. First, a neutral formant post-filter is constructed for the current frame, and the delay line of the filter is initialized with zeros.

$$H_{neutr}(z) = \frac{1 - 2 \cdot 0.9cos(\theta_1)z^{-1} + 0.9^2 z^{-2}}{1 - 2 \cdot 0.9cos(\theta_1)z^{-1} + 0.9^2 z^{-2}} \cdot \frac{1 - 2 \cdot 0.9cos(\theta_2)z^{-1} + 0.9^2 z^{-2}}{1 - 2 \cdot 0.9cos(\theta_2)z^{-1} + 0.9^2 z^{-2}} \tag{3.9}$$

Then, the interpolation is done between the neutral post-filter of Equation (3.9) and the actual post-filter with the correct values for r_1 and r_2 of the next frame. This operation has the advantage of both smoothing the transition as well as updating the memory of the post-filter to reasonable values. If there is no smoothing period, and the filtering is just suddenly turned on, the filter memory contains random values or zeros, and this sometimes causes sudden energy peaks in the speech signal which often result in audible artifacts.

AGC

The adaptive gain control algorithm is used to adjust the energy of the processed signal to correspond to that of the original speech signal. The speech frames are processed in 5 ms subframes with the same algorithm that is used in the AMR narrowband standard [10]. First, a scaling factor is computed by using the following equation.

$$\gamma = \sqrt{\frac{\sum_{n=0}^{39}(s(n))^2}{\sum_{n=0}^{39}(s_{pf}(n))^2}} \tag{3.10}$$

In Equation (3.10), $s(n)$ is the received signal and $s_{pf}(n)$ is the signal that has been post-filtered. The output of the algorithm is given as

$$s_{scaled}(n) = \beta(n)s_{pf}(n), \text{ where} \tag{3.11}$$

$$\beta(n) = 0.9 \cdot \beta(n-1) + 0.1 \cdot \gamma \tag{3.12}$$

The values of $\beta(n)$ are calculated for every sample, and are used to smooth the changes between samples.

3.3 SII calculation

The speech intelligibility index of the speech sample is determined just before the noise is added. The functions that are used for this purpose were obtained from the SII website [38] which has several implementations that follow the requirements defined in the standard [26]. The MATLAB implementation was slightly modified before use in order to make the processing faster.

The speech level of the samples is assumed to be 70 dB SPL (Sound Pressure Level) which is a normal conversational level. The noise level is thus $70 - SNR$. Both the information signal and the noise are converted using a function from the HUTear toolbox [39]. This has to be done because the signals are scaled in MATLAB to lie between +1 and -1. In order to be able to estimate the intelligibility value, their true levels have to be determined.

Before the intelligibility index is calculated, the silent periods are removed from the calibrated speech sample. If this phase was omitted, the SII value would be inaccurate because the silent periods unnecessarily lower the mean value. Even though these time regions contain absolutely no speech information, they contribute to the mean value of the sample. Their intelligibility is zero because they are completely masked by the noise.

The speech intelligibility is calculated in 9.4 ms frames. The actual SII value given to a particular speech sample is the mean intelligibility over all of the frames. The spectrum levels for both the speech and noise are calculated in 18 frequency bands which are called the *one-third octave bands*. After this, the proportion of speech information available in each band is determined and given as a number between zero and one. All of the 18 bands are weighted with a band importance function which can be chosen from seven different possibilities defined in the standard. These include functions for speech materials which consist of nonsense syllables or phonetically balanced words, for instance. The option closest to the available speech material was defined as short passages of easy reading material, and thus this function is used to weight the frequency bands.

Chapter 4

Subjective tests

The purpose of this chapter is to discuss the subjective listening tests that were conducted. Their motivation was to search for suitable parameter values for the post-processing algorithm. The two parameters to be determined were the denominator parameters of the formant post-filter, r_1 and r_2. First, the methods that were used are introduced in detail, and after this, the results are presented and briefly discussed. Finally, the results are analyzed in terms of the speech intelligibility index.

4.1 Methods

The main goal of the subjective listening tests was to determine suitable parameter values for the post-filter instead of evaluating the speech quality given by the algorithm. Because there are no standards covering this kind of situation the test had to be designed from scratch. Questions, such as whether the focus would be on intelligibility or quality and how the test samples should be chosen, had to be answered in the process.

In the end, the focus of the test was neither on quality or intelligibility, but somewhere in between. In a formal intelligibility test, each sample is usually heard only once. If the sentence remains the same, it is hard to overlook a possible learning effect in the results. This means that once the listener understands the contents of a sentence, even the samples that were unintelligible before are heard correctly. In this case, since the post-filter shifts energy to the higher frequencies and thus makes the speech clearer, the learning effect is a real problem. The neutral reference may be hard to understand at first, but after listening to some processed samples that are more intelligible, the reference is also heard correctly. However, the question was about listener preferences. Even though the written instructions given to the listeners guided towards considering all of clarity, quality and naturalness, it was not guaranteed that the neutral reference would not be preferred.

Another big question was the type of user interface and the amount of different processing conditions presented to the listeners. Usually in formal listening tests, the samples are pre-processed and the listener has a finite amount of them to grade or to choose from. Here one possible approach could be to form a grid in the two-dimensional parameter space. Depending on the spacing of the parameters, the listeners would have M different samples to listen to, and they would be asked to choose the best one. The spacing of the grid would in this case have a large effect on the results. With a very sparse parameter grid, the results would probably show smaller deviation, but the subjects might never hear the sample that they would otherwise consider the best. With a very frequent spacing, there would be too many samples that the listeners would have to listen to. In the end, it was decided that the test subjects could freely choose the parameter values from the given ranges, $0.27 \leq r_1 \leq 0.9$ and $0.9 \leq r_2 \leq 0.99$. This approach enables the test subjects to form their own grids, based on their preferences and on their ability to hear small details and differences. On the other hand, the results would probably have large deviation.

The graphical user interface that was used in the test is depicted Figure 4.1. It consists of two push buttons, one marked as neutral and the other as next, and a blank, white space with two axes. The idea is that by clicking the neutral button the listener can play an unprocessed version of the current sample. By clicking somewhere in the white area, the sample processed according to the coordinates is played. The processing of the samples is done in real-time. This, of course, requires that the delay is very small or otherwise the listeners would be annoyed by the waiting time. After the processed sample has been played, a red square will appear on the spot that was clicked on to mark the location. Samples can be listened to again by clicking on the squares that have appeared on the screen. The red color always denotes the sample that was last heard while the other markers are blue.

In the user interface, the x-axis corresponds to the parameter of the first formant, r_1, and the y-axis to that of the second formant, r_2. By moving further away from the neutral point, the processing naturally becomes more extreme. This means, for example, that in the lower right corner, the attenuation of the first formant is at maximum. However, the test subjects were only told that some kind of processing was done and that its effects would grow more drastic linearly as the distance from the neutral corner would be increased.

Speech material

Six speech samples from six different speakers were used in the subjective test. Three of the speakers were male and three female. The material was in Finnish and the speakers were native in the language. The samples chosen for the test came from a set of high quality recordings where each of the speakers was asked to read the same written text which dealt with weather forecasts. For the subjective test, a different short sentence was chosen from

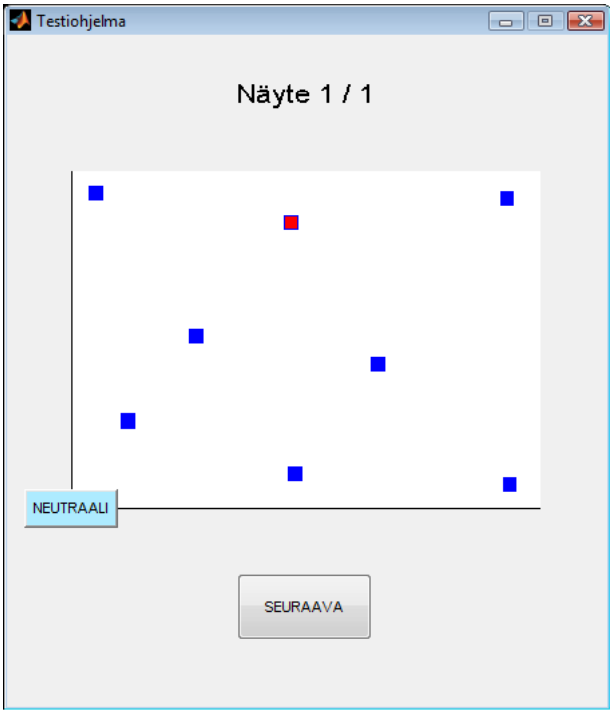


Figure 4.1 – The interface used in testing. The square markers on the screen have been added by the listener.

the material for each speaker in order to cover larger phonetic variance. The test sentences along with their speakers are contained in Appendix B. In the test the order of the speakers was randomized.

Car noise was added to the test samples, so that the resulting SNR of the sentences was -5 dB. The calculation of the signal-to-noise ratio was conducted with the method explained in Chapter 3. The noise level was chosen on purpose to be very high, and most of the unprocessed samples were difficult to understand completely the first time they were heard.

Listeners

A total of 18 test subjects took part in the subjective listening test. Seven of them were female and 11 male, and their ages were between 21 and 45 with an average of 26.8 years. All of the listeners were naive according to their own evaluation, and 14 of them had a technical background. The participants were all required to speak and understand Finnish. The test subjects were not paid for their participation.

Test situation

The listening test was conducted in a quiet office space with Sennheiser HDA 200 headphones which were chosen for the purpose because of their effective insulation. Before the actual test, the subjects were given a short instruction paper to read through. The original Finnish text is included in Appendix C. It briefly explains how the interface works and what the listener is expected to do in the test. After the instructions were read, a one sample demo was available. The idea of the demo was to familiarize the subjects with the graphical user interface and to also choose an appropriate volume level for the test. After and during the demo the listeners were allowed to ask questions, and the experimenter could also observe whether they seemed to grasp the idea of the test or if they needed further guidance.

The six sample test took between 20 and 45 minutes depending on the subject. During this time the supervisor of the test was also able to discreetly observe the listeners' actions, and after the test some questions were asked about the test and their observations on what happened to the samples because of the processing. For each of the six samples, all of the locations of the subject's clicks on the screen were stored. One purpose of all of these measures was to make sure that the test subjects had understood the idea of the test correctly, and also to weed out possible accidental erroneous choices. The listeners were told that the red square would also mark their final choice when moving on to the next sample, but there were no guarantees that they would always remember this. There was also a distinct possibility that the inherent logic of the two-dimensional parameter space would not be understood. In this case, the reliability of the subject's results would be under question.

4.2 Results and discussion

Out of the 18 test subjects only one had to be discarded from the final results. By looking at the subject's data and the answers that were given to the questions after the test, it became evident that the subject failed to understand the point of the test. The values that were chosen were actually not entirely different from the rest, but their motivation was incorrect. An example of the data from the tests from one of the accepted subjects can be seen in Appendix D. Most of the listeners had a very similar approach to the test. At first, they formed some kind of grid, and after locating a good area, the grid was tightened in that region.

The final results from the 17 other listeners are presented in Figures 4.2 and 4.3 by sample and by speaker gender respectively. In Figure 4.2, the first three speakers (HaPu, MaAi and PaAl) are male and the final three (HeLe, LaLe and VeAl) are female. In both figures the final choices of the listeners are presented with red markers and the mean of the values with a blue marker. The standard deviations in both directions are also included. The exact mean

values for the parameters are also given in the titles after the speaker identities. The grey circles are outliers that were not taken into account when calculating the mean values. They were deemed to be outliers because the two processed samples have a strong whistling effect and are therefore extremely irritating to listen to. It was concluded that they were marked as final choices by accident.

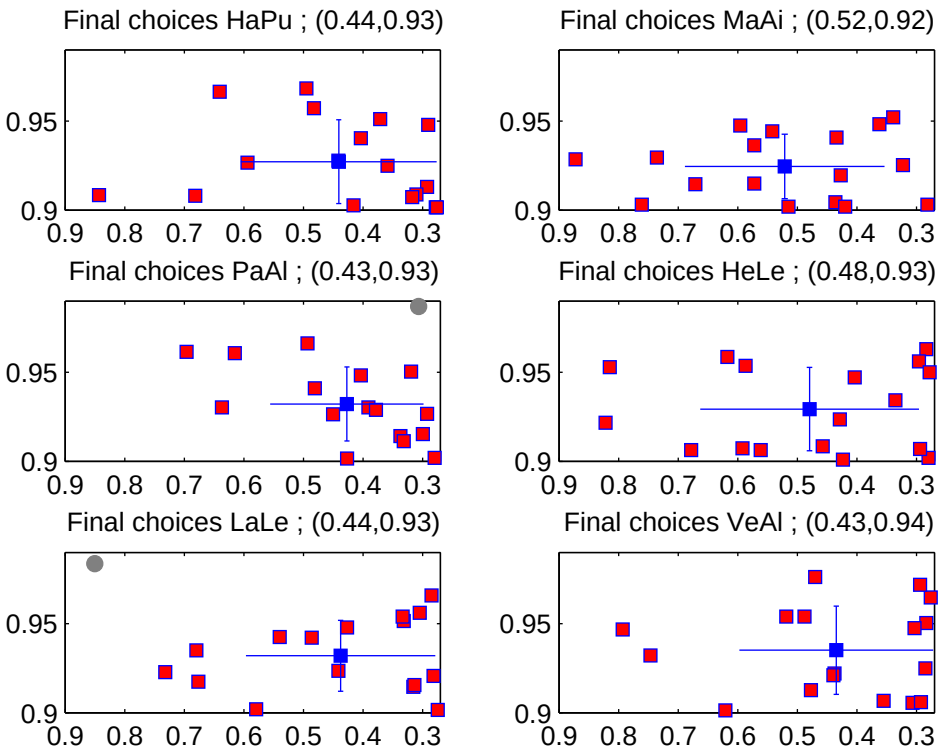


Figure 4.2 – Test results by speaker. The red square markers denote the final choices of the listeners, the blue square markers are the mean values and the grey circles are outliers.

The preferences of the listeners vary greatly as was expected, but a general tendency towards the right side of the area can be seen in both Figures 4.2 and 4.3. Most of the listeners characterized the right lower corner as being the clearest and that the speech was there easier to separate from the noise. Many commented that the test sentences gave them a distinctive news-like feeling and as a result they felt that clarity was the most important thing to consider. On the other hand, a few listeners said that the naturalness of the voice was compromized and that the processing on the x-axis made the speaker sound urgent and nervous. These listeners tended to prefer a very neutral processing that was found in the

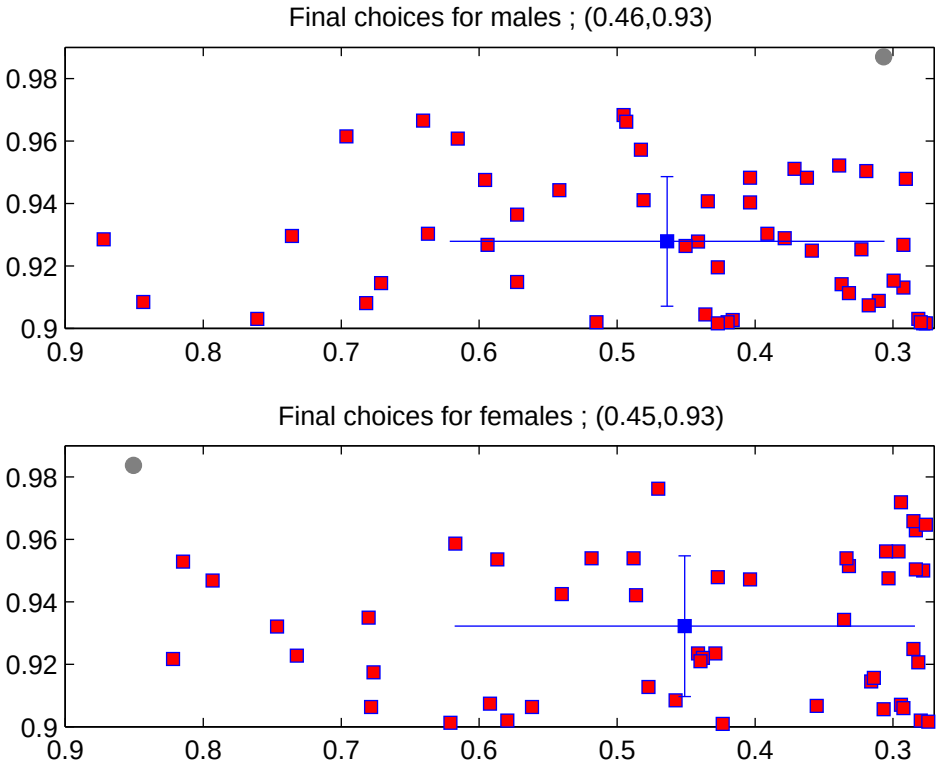


Figure 4.3 – Test results by speaker gender. The red square markers denote the final choices of the listeners, the blue square markers are the mean values and the grey circles are outliers.

lower left corner.

Another common observation among the test subjects was that a high value on the y-axis resulted in irritating speech with poor quality. It was characterized, for example, as unpleasant, distorted and metallic. There can also be heard a distinctive whistling effect which is probably due to the fact that the second formant peak becomes extremely sharp when r_2 approaches 0.99. Whistling sounds are characterized by a very narrow peak with high amplitude in the 500 to 3000 Hz frequency range [40]. On the other hand, some test subjects who preferred very neutral processing reported that some change in the vertical direction was less irritable than in the horizontal direction because it did not color the speech in the same way.

Most of the listeners felt that different samples were affected by the processing in a very similar way. Although, as can be seen from Figure 4.2, there a some differences between the optimum parameters. Some test subjects noted that the female voices were somehow

more understandable and easier to separate from the noise even in the unprocessed sample, and that the processing did not affect them as dramatically as the male voices. The reason behind this could be that female speakers tend to have higher formant frequencies than male speakers [41]. This means that a larger proportion of speech information is already available before the processing. The effects of the post-processing could also be diminished because the first formant is higher, and thus the energy is not necessarily moved from the frequency band where most of the noise energy is concentrated, but from some frequency region above that.

In this light, it is interesting to see that the mean parameter values are almost the same for males and females in Figure 4.3. However, the male-female categorization is rather crude because there can be male speakers with high fundamental frequency and formant frequencies as well as female speakers with a low F0. It was also quickly tested, whether the mean values would correlate with the fundamental frequencies of the speakers. The results indicated that there is no correlation between the two, but it should be remembered that there is only a small amount of data to test and it has a large variance. It would be interesting to see, what would happen, if a larger amount of more controlled results could be analyzed. Based on the test results, the parameters chosen for the formant post-filter were $r_1 = 0.46$ and $r_2 = 0.93$.

4.3 Speech Intelligibility Index

An additional interest was the behavior of the speech intelligibility index both in general and compared to the subjective decisions. Figure 4.4 presents the speech intelligibility index contours and the final choices of the test subjects for the male speakers. Figure 4.5 contains the same information for the female speakers. Once again, the grey circles mark the samples that were removed as outliers. The numbers associated with the curves in the figures tell that the value of the speech intelligibility index inside the contours is higher than this limit. For example, here the samples to the right of a curve have an SII value at least as high.

The first observation that can be made based on these figures is that the value of the speech intelligibility index seems to grow when moving towards the lower right corner. This was to be expected because in that direction the first formant is attenuated and an increasing portion of the energy is shifted to higher frequencies, thus making the speech stand out better from the noise in certain frequency regions. It is also interesting to note that, according to the contour curves, the enhancement of the second formant actually reduces intelligibility after some point. For example, with speaker PaAl the curves start out as leaning towards the left which means that the SII value is increased by moving in the

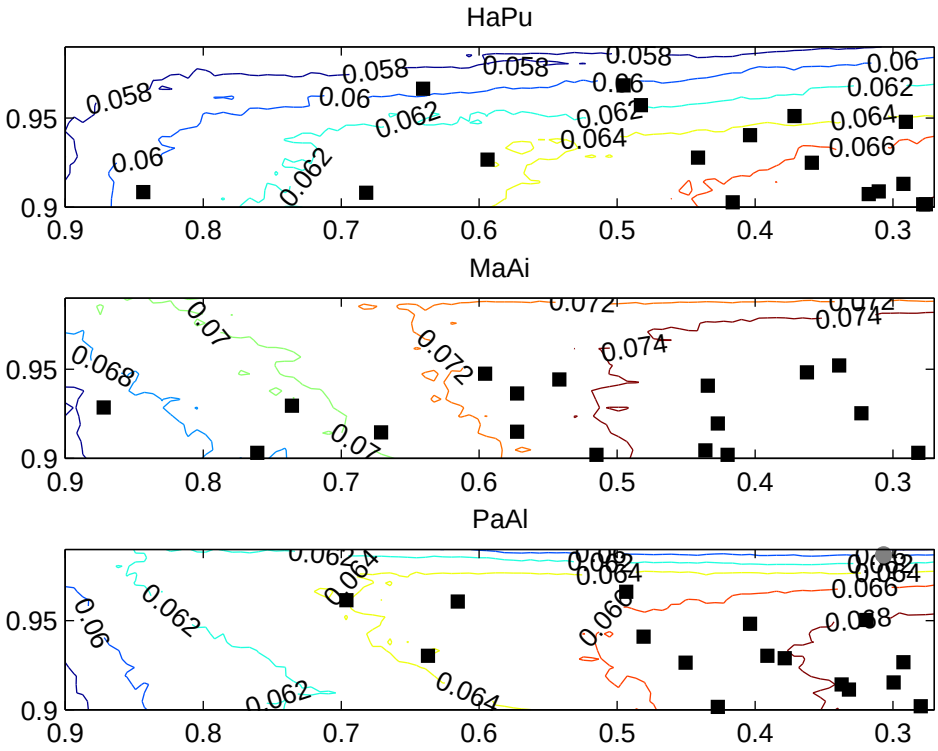


Figure 4.4 – The SII contours for the male speakers along with the results from the subjective test.

direction of either axis. However, when the value on the x-axis goes below 0.7, the contour curves start excluding some of the highest values on the y-axis. This behavior could be explained by assuming that in order to have a very sharp peak energy has to be taken from some other frequency band in the spectrum because the overall energy level is scaled after the processing. After some point, this energy increase stops benefiting the SII because the frequency band where the energy is added to has reached its maximum contribution. At the same time, other frequency regions which could potentially increase intelligibility are deprived of energy thus further reducing their speech intelligibility value. This assumption is also supported by the fact that in most of the figures the curves only start presenting this behavior after the horizontal parameter has been moved. When the first formant is attenuated, the energy is shifted and some of it lands on the frequency band of the second formant. In this case, the sharpening of the second formant has an even smaller effect.

The values related with the speech intelligibility index contours are very small. As was mentioned earlier in Chapter 2, the possible range of SII values is from zero to one, where

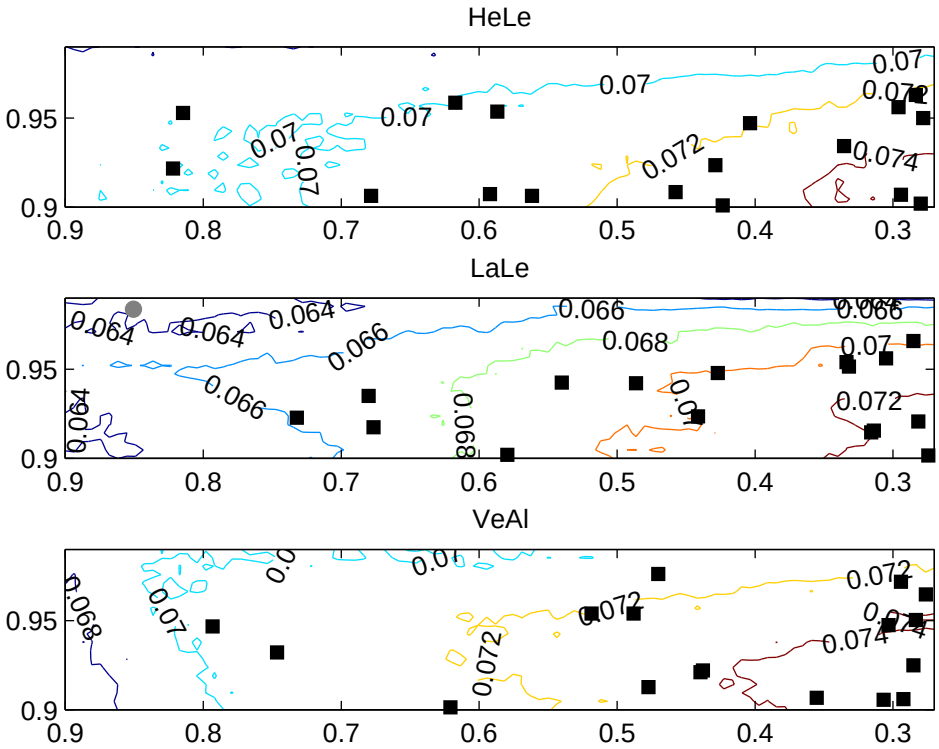


Figure 4.5 – The SII contours for the female speakers along with the results from the subjective test.

zero is completely unintelligible and one denotes perfect intelligibility. In this case, the values range only from 0.058 to 0.074. This tells that the level of the degrading noise was indeed very high and that the unprocessed speech was hard to understand. The test subjects had trouble understanding the unprocessed samples for the first time. However, all of them did understand the samples completely at least with some of the parameter values. This raises a question on the quality of the SII measure. How is it possible that a sentence is understood perfectly when according to the calculated SII value one would imagine that the speech should be almost completely unintelligible?

When looking at the final choices of the listeners which are marked with black squares, it is hard to see a clear correlation with the speech intelligibility index contours. Of course, most of the markers are concentrated in the lower right corner, but they do not seem to follow any pattern and some of the markers are more spread out. On the other hand, the differences between the SII values on the figures are small compared to the whole range of the index. It is possible that they are completely inaudible in which case the listeners'

choices could not be expected to follow them.

Even though the noise was stationary as required in the standard defining the SII, it is still unclear whether this type of measurement is suited for this purpose. Perhaps some sort of re-calibration of the values should have been done in order to receive more valid and valuable data. Furthermore, choice of the band-importance function has a large effect on the form of the contours and the magnitude of the values when the differences are so small.

It is impossible to know what kind of results should be given by the SII calculation without conducting a subjective intelligibility test. These results could be used to investigate the effects of the band-importance function and to determine the most suitable one. An entirely different problem is the effect of language on the speech intelligibility index measure. There is a possibility that it only gives reasonable results when the material is in English.

Chapter 5

Objective evaluation

The purpose of this chapter is to evaluate some charateristics and behavior of the final post-filter. Since any type of formal listening tests to assess the possible quality or intelligibility improvements of the filter scheme were not yet conducted, the focus will be on more objective measures. First, the gains of the general post-filter structure and the effects of the parameters, r_1 and r_2, on them will be discussed. After this, its effects on typical voiced sounds are presented. Finally, some comparisons between the developed post-filter and two other post-filtering approaches from Chapter 2 will be made.

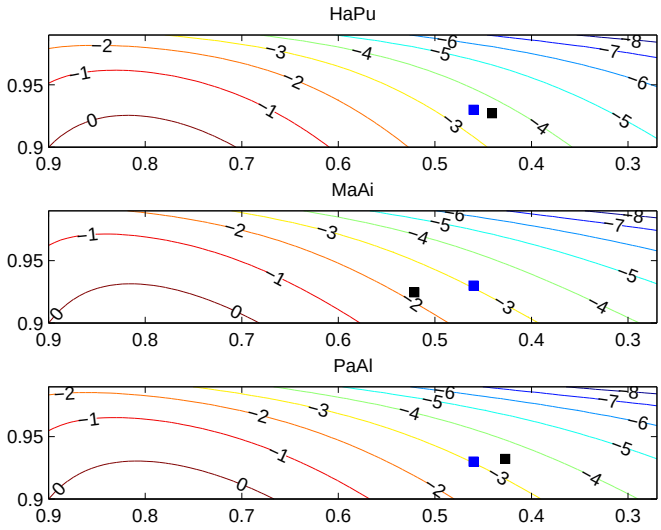
5.1 Post-filter gains

The decibel gains of the post-filter were evaluated using Equations (5.1) and (5.2) which were derived in Appendix A.

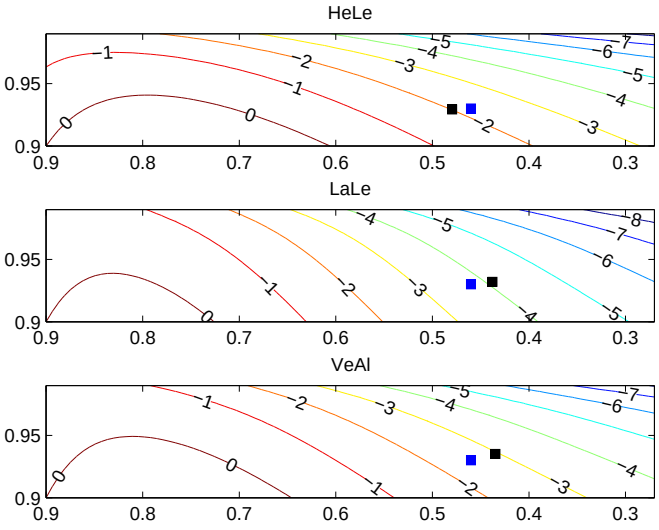
$$\left|H(e^{j\omega})\right|^2 = \frac{f(0.9,\theta_1;\omega)}{f(r_1,\theta_1;\omega)} \cdot \frac{f(0.9,\theta_2;\omega)}{f(r_2,\theta_2;\omega)} \cdot \frac{1}{1-2\mu cos(\omega)+\mu^2} \quad \text{, where} \tag{5.1}$$

$$f(r,\theta;\omega) = 1 + r^4 + 4r^2cos^2(\theta) - 4rcos(\theta)cos(\omega)(r^2+1) + 2r^2cos(2\omega) \tag{5.2}$$

In the equations, parameters θ_1 and θ_2 are the locations of the first two formants, r_1 and r_2 the filtering constants and μ is derived from the overall tilt of the filter. The angle ω denotes the frequency where the gain is calculated. The parameter r_1 was allowed to vary between 0.27 and 0.9 and r_2 between 0.9 and 0.99, and the decibel gains were calculated at four different frequencies over the samples used in the subjective test. These frequencies were 0 Hz and 4000 Hz as well as the frequencies of the first and the second formants. Of course, the locations of the formants change constantly as the shape of the vocal tract changes which also affects the gains. For this reason, the gains were calculated for each voiced frame and then averaged, so that the results are the average attenuation of the first formant in decibels and the average enhancement of the second formant in decibels. The contour plots based



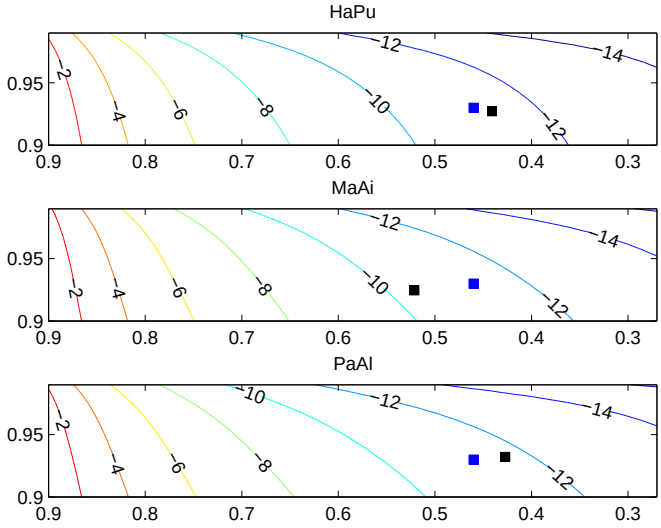
(a) Male speakers



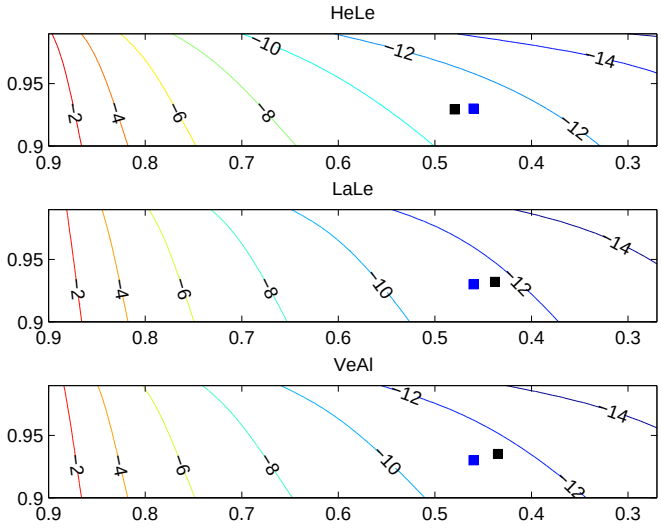
(b) Female speakers

Figure 5.1 – The average gains at 0 Hz along with the means from the subjective tests and the chosen parameter values.

on these data can be seen in Figures 5.1(a)- 5.4(b). In each figure, the mean parameter value for that particular speaker obtained from the subjective tests is marked with a black square. The blue markers denote the chosen post-filter parameter values. The contour labels refer



(a) Male speakers

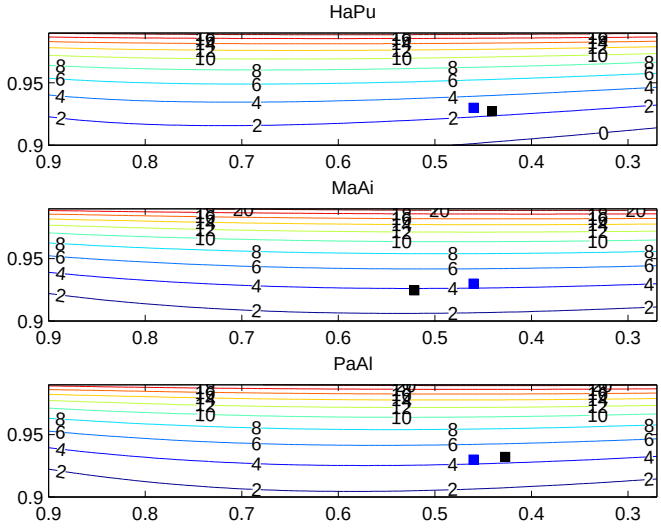


(b) Female speakers

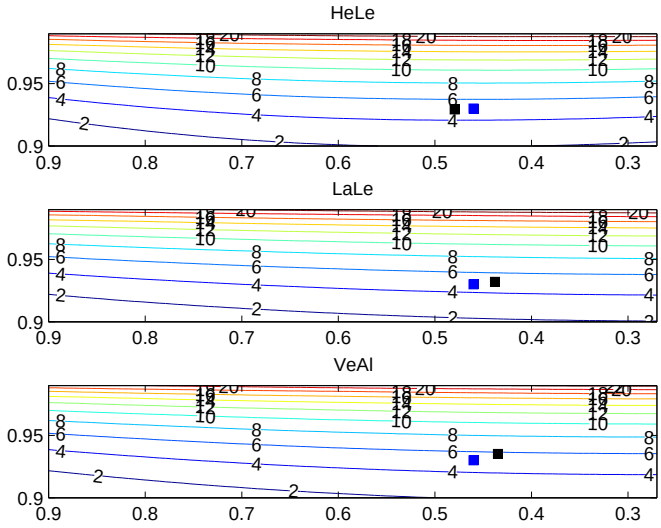
Figure 5.2 – The average gains at the first formant frequency along with the means from the subjective tests and the chosen parameter values.

to the values inside the contour. In other words, the gains inside the curves are larger than the values on the border.

The contour plots are very similar between the samples in all cases. Of course, the



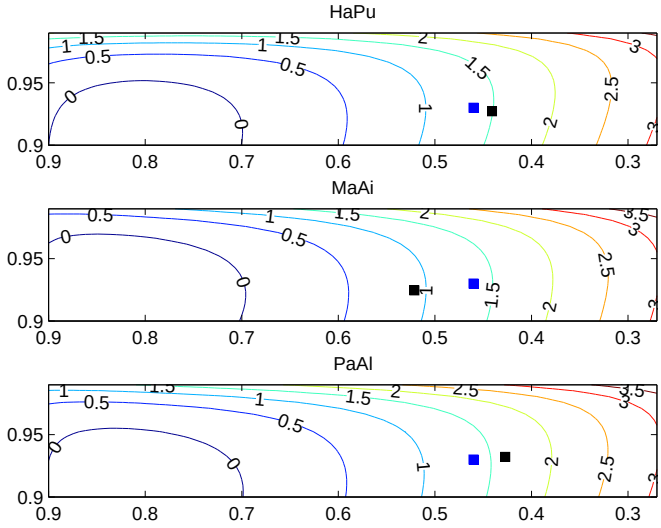
(a) Male speakers



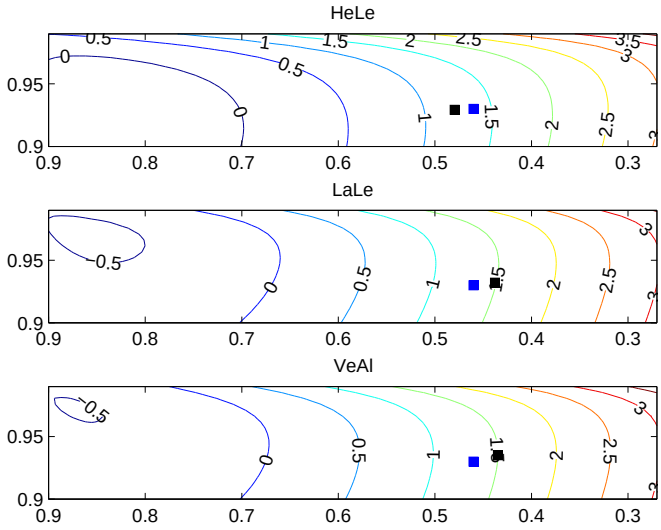
(b) Female speakers

Figure 5.3 – The average gains at the second formant frequency along with the means from the subjective tests and the chosen parameter values.

fact that they are averaged over the whole sample further diminishes possible differences between speakers. The gains at 0 Hz, depicted in Figures 5.1(a) and 5.1(b), are mostly quite small, around -3 dB for the means of the samples as well as the chosen post-filter



(a) Male speakers



(b) Female speakers

Figure 5.4 – The average gains at 4000 Hz along with the means from the subjective tests and the chosen parameter values.

parameters. As the attenuation of the first formant continues to increase, the amplitude level at 0 Hz also begins to drop because the frequency locations are so close to one another. The second parameter, r_2, also affects the gain at 0 Hz, but in a different way. As the peak of

the second formant becomes sharper, the gain reduces.

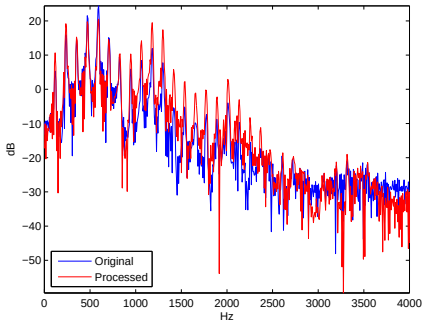
The average attenuations for the estimated first formants presented in Figures 5.2(a) and 5.2(b) reach much larger values as was expected. The gains for the means from the subjective tests and for the chosen post-filter values are around -11 dB. The change in parameter r_1 has a larger influence on the gain which is only natural because it affects the first formant directly. As r_1 becomes smaller, the dip on the first formant frequency becomes larger.

Figure 5.3 presents the contour plots for the gains at the second formant frequency. The first thing to notice is that the closer the pole gets to the unit circle in the z-plane, the closer the contour lines get to each other. In other words, when r_2 approaches 0.99, even a small change in the value has a large impact on the decibel gain. At the top of the figures, the average gains reach 20 dB which means that the formant peak has to be extremely sharp. The strong whistling effect which was earlier discussed in Chapter 4 is probably related with this. However, the mean values from the subjective tests have gains of only around 4 to 6 dB.

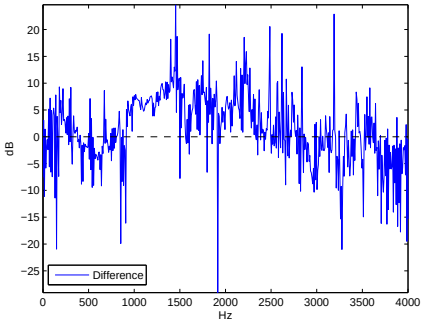
The average gains at 4000 Hz are given in Figure 5.4. They are once again very small, and this is due to the fact that neither of the formant frequencies is nearby, so their effect is negligible. The gains for the mean values and for the chosen parameter values are around 1.5 dB.

5.2 Typical behavior

In addition to analyzing the gains of the proposed post-filter, its effects on some typical voiced sounds were studied. Figure 5.5 shows the vowel sound [a] uttered by a male speaker. The image on the left displays the amplitude spectrum of the vowel before and after processing with the post-filtering algorithm. In order to highlight the main differences between the amplitude spectra of the two signals, the figure on the right contains the difference between the amplitude spectra of the processed and the original signals. The difference of the amplitudes of the two signals is not necessarily the best way to display information because it tends to have large spikes, but it does give some overall tendencies. Figure 5.6 presents the same images for the liquid [l] uttered by the same male speaker. For comparison, in Figures 5.7 and 5.8 are the same voiced sounds, but this time spoken by a female speaker. The approximate formant frequencies as estimated by the algorithm, and the gains of the post-filter for these phones are given in Tables 5.1 and 5.2. All of these examples were extracted from whole words that had first been processed with the post-filtering scheme. This means that they are also AMR coded, but no noise has been added to the processed signals.

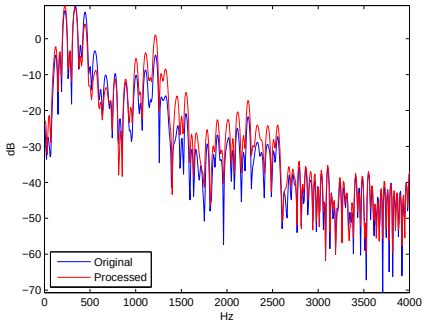


(a) The spectra of the original and the processed signals.

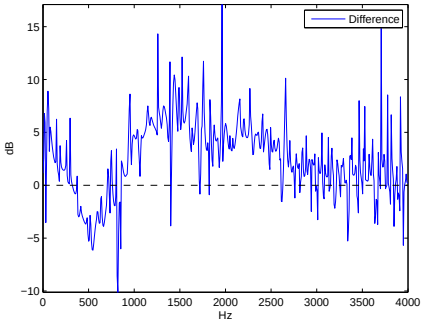


(b) The difference: $s_{proc} - s_{orig}$.

Figure 5.5 – The effects of the processing on the vowel [a] for a male speaker.



(a) The spectra of the original and the processed signals.



(b) The difference : $s_{proc} - s_{orig}$.

Figure 5.6 – The effects of the processing on the liquid [l] for a male speaker.

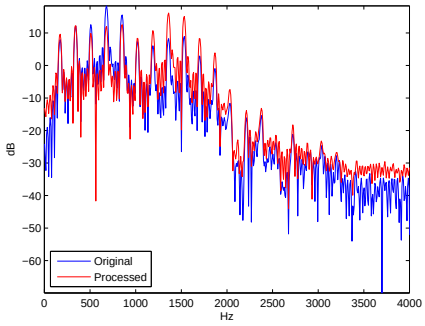
Table 5.1 – The average gains for typical voiced sounds for male speakers. The frequencies denoted with a $\sim$ are the estimated first and second formant frequencies respectively.

(a) The vowel [a].

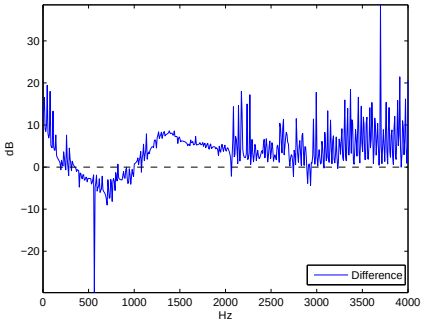
Frequency	Gain
0 Hz	-1.5 dB
$\sim$594 Hz	-10.9 dB
$\sim$1375 Hz	4.4 dB
4000 Hz	1.3 dB

(b) The liquid [l].

Frequency	Gain
0 Hz	-2.2 dB
$\sim$562 Hz	-11.0 dB
$\sim$1437 Hz	4.8 dB
4000 Hz	1.4 dB

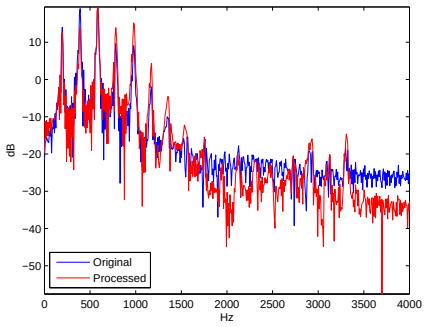


(a) The spectra of the original and the processed signals.

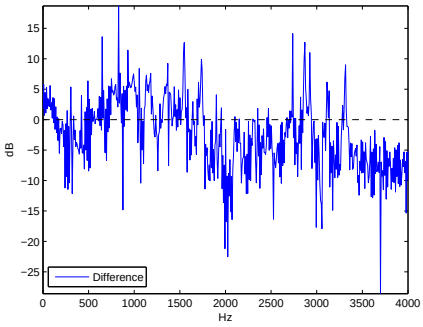


(b) The difference: $s_{proc} - s_{orig}$.

Figure 5.7 – The effects of the processing on the vowel [a] for a female speaker.



(a) The spectra of the original and the processed signals.



(b) The difference : $s_{proc} - s_{orig}$.

Figure 5.8 – The effects of the processing on the liquid [l] for a female speaker.

Table 5.2 – The average gains for typical voiced sounds for female speakers. The frequencies denoted with a $\sim$ are the estimated first and second formant frequencies respectively.

(a) The vowel [a].

Frequency	Gain
0 Hz	0.4 dB
$\sim$687 Hz	-10.5 dB
$\sim$1344 Hz	3.6 dB
4000 Hz	1.2 dB

(b) The liquid [l].

Frequency	Gain
0 Hz	-6.7 dB
$\sim$406 Hz	-12.4 dB
$\sim$1625 Hz	5.6 dB
4000 Hz	1.5 dB

Based on the figures, all of these phones seem to be affected in an almost identical way by the processing. The two signal spectra, the reference and the processed one, look very similar to each other in all cases, but some of the main differences can be picked out from the figures on the right side. Around 500 Hz, there is a frequency region where the difference of amplitudes is negative. In other words, the amplitudes of the original signals are higher than those of the processed ones in this region. After around 1000 Hz, the difference of amplitudes reaches positive values and they stay that way until approximately 2500 Hz. This was to be expected as the idea is to move some energy from low frequencies to higher frequencies. Because the value of r_2 is only 0.93, the second formant is not significantly sharper than the rest of the peaks.

In some cases, the difference of amplitudes has large positive values below 250 Hz. In other words, some of the energy from the first formant frequency has been moved to an even lower frequency, and therefore the processed signal has higher amplitude than the original signal in that frequency region. This is not a desirable phenomenon since the idea was to shift energy to higher frequencies where the energy level of the noise is lower. In Figures 5.6 and 5.7 the effect is evident.

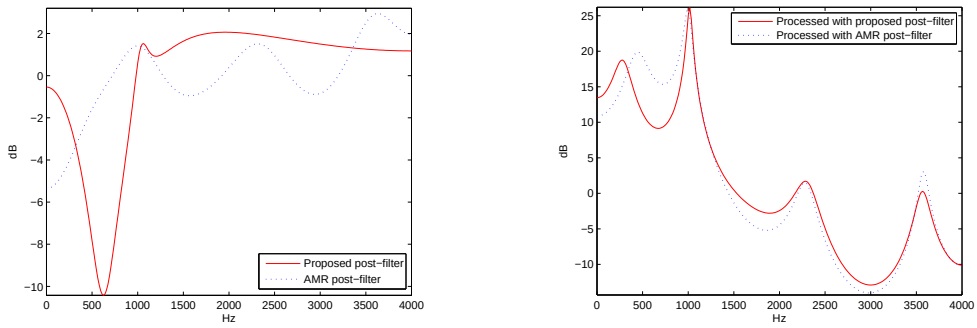
In Figure 5.8, the reference signal has higher amplitude values than the processed signal at high frequencies. Whereas in the other cases the difference of the amplitudes is around zero near 4000 Hz, here the values are clearly negative. This means that the fourth formant is not enhanced, but in fact attenuated. It could be caused by the tilt compensation which actually tries to prevent the post-filter from enhancing the fourth formant too much. The question is, why is the phone [l] uttered by a female speaker affected more clearly than that given by the male speaker? Of course, it should not be forgotten that the characteristics of a phone are also affected by its surroundings. In other words, the spectrum of the liquid [l] looks different when it is extracted from the Finnish word ”saatavilla” instead of ”avulla”.

The gains presented in Tables 5.1 and 5.2 are close to each other in all cases. The estimated formant frequencies are the two middle rows in the tables. The estimated frequencies for the vowel phones [a] are similar to standard values in Finnish so the estimation is known to be somewhat correct. One thing that jumps out from the tables is the fact that the liquid [l] uttered by a female speaker also has the widest gap between the two formants. As discussed earlier, it displayed some unexpected behavior at frequencies near 4000 Hz. Perhaps this also contributes to the phenomenon.

5.3 Comparison with other post-filters

Naturally, it is interesting to see how the proposed post-filter compares to the other post-filtering structures that were discussed in Chapter 2. The most obvious choice is to compare

it to the one in the AMR narrowband standard which is a logical reference scheme because it is the current implementation in the standard. This comparison is presented in Figure 5.9 which contains the amplitude responses of the two filters on the left side. On the right side, there are two LP spectra, one for each filter. A vowel sound, [a], has first been processed with both filters, and then the LP spectra has been calculated from the processed signals.



(a) The amplitude responses of the filters.

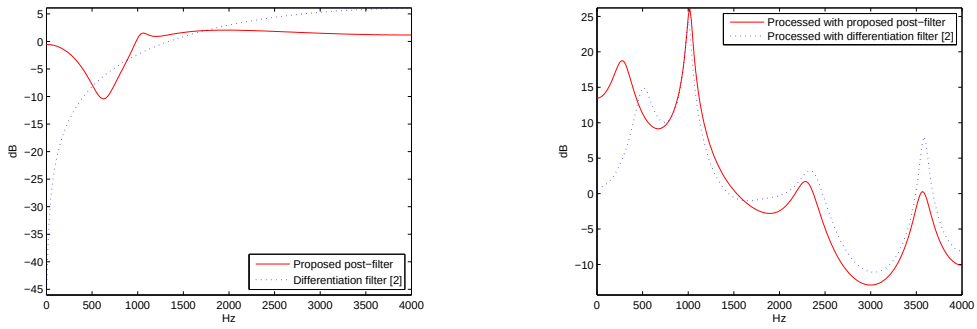
(b) The effects of the filters in LP spectra.

Figure 5.9 – The proposed post-filter and the AMR post-filter [10].

The filters differ to a great extent based on Figure 5.9(a). The standard post-filter has small gains throughout the spectrum and does not modify the speech signal very much. In contrast, the designed post-filter has larger gains, and consequently, the speech signal goes through more dramatic changes. The reason for this difference is that the filters have two distinctively different design approaches. Whereas the proposed post-filter is meant to work in severe noise conditions where the signal can be modified more freely, the standard AMR post-filter tries to avoid changing the signal very much in fear of lowering the audible quality of the processed speech. Under these circumstances, an objective comparison between the two is hard to make. However, one noticeable difference is that the AMR post-filter tries to attenuate the valleys between formants, but the developed post-filter actually enhances some of them.

Also another post-filter is used for comparison, namely the differentiation filter by Hall et al. [8] which was presented in Chapter 2. This filter was chosen because the basic idea is the same in both cases even though the intended application scenarios are a little different. Their mutual goal is to move more of the energy to higher frequencies using extreme measures. The post-filters and their effects on the LP-spectra of phone [a] are depicted in Figure 5.10.

The differentiation filter is clearly a simple high-pass filter as can be seen from Figure 5.10(a), and as such requires much less computation than the proposed post-filtering scheme. It attenuates the lowest frequencies much more drastically than the developed post-filter. On the other hand, the higher frequencies naturally have a strong tilt, which



(a) The amplitude responses of the filters.

(b) The effects of the filters in LP spectra.

Figure 5.10 – The proposed post-filter and the differentiation filter by Hall et al. [8].

is not a desired effect. The proposed post-filter has a much flatter frequency response at the highest frequency band which helps in making the speech sound more natural. The strong high-pass effect of the difference filter is also seen in the LP spectra where the fourth formant is enhanced so that its amplitude becomes larger than that of the third formant. The differentiation filter also enhances the valleys between some formants, but the effect is stronger than with the developed post-filter.

One large issue which clearly shows in the previous figures is that the proposed post-filter moves the first formant to a lower frequency band as mentioned earlier. It is unclear as to how this affects the quality of the processed speech signal, if at all, and whether this phenomenon happens with all speakers and voiced phones, or is it restricted to some subset of them. Previously, it was noted that the effect was more evident in some phones than in others. Part of the problem could be the estimation of the formants, even though the estimated frequencies earlier seemed to fit the standard frequency ranges. This does not mean that the estimates are always accurate. Also the post-filter structure itself is problematic since it enhances or attenuates some other frequency regions close to the actual formants, in spite of having the correct values of θ_i. The difference in frequency is supposedly small, but the actual analytical solution has not been calculated.

Chapter 6

Conclusion

6.1 The contribution of this work

In this thesis, a new post-processing algorithm for speech enhancement has been proposed. It is designed to work with speech in difficult noise conditions where the intelligibility of the speech is seriously compromised. This means that the level of the noise is so high that it makes the speech hard to understand. The situation is that the listener is in the receiving end of a mobile phone connection, and the post-processing happens at the receiver's mobile device. The only noise source is environmental noise at the receiver's end. In other words, the speech received from the channel is assumed to be noiseless, and the post-processing is done before the degrading noise has affected the speech signal.

The post-filter works by shifting energy from a low frequency band, where the noise is the strongest, to the higher frequencies, thus increasing the amount of speech information available to the listener. The reason is that at higher frequency bands, the speech is no longer masked by the noise. First, the algorithm picks out the voiced frames from speech, and estimates the locations of the formants in them. Then, the first formant is attenuated by approximately 10 dB, and the second formant is enhanced by approximately 5 dB. The tilt of the formant filter is used to construct another filter with the purpose of preventing an undesirable high-pass effect in the processed speech. After the filtering, the energy of the processed signal is equalized to the energy level of the original speech signal.

Two parameters of the post-filter were obtained by conducting an informal subjective listening test with 18 test subjects. In the test, the listeners were asked to locate their preferred processing for the samples in a two-dimensional space. The mean values from their choices were adopted to the final post-filter structure. After this, the behavior of the developed post-filter was studied with different objective methods. In other words, it was investigated how the post-processing affected common voiced sounds spoken both by male

and female subjects. The post-filter was also compared to some of the post-filter realizations that have been previously used by other authors.

Based on these evaluations, it was concluded that the post-filter works in the desired way for the most part. It also presented some unexpected and unwanted behavior which needs to be further studied to determine its causes and the resulting audible effects on processed speech. It is difficult to conclude how good the post-filter actually is when there are no performance results from formal listening tests. However, the proposed post-filter manages to bring a new, more adaptive method to post-processing in high noise levels. Previously, this problem has been approached with simple, static filter structures that improve the intelligibility to the detriment of quality. With the developed post-filter it is possible to fine tune the processing, so that the quality of the speech does not suffer.

6.2 Practical implementation

Since the ultimate goal is to develop a post-filtering scheme working in real-time in a mobile phone, the requirements and specifics for the practical implementation should also be discussed. For now, everything has been done using MATLAB, and the situation that would be in an actual mobile device has been simulated carefully. However, some unrealistic assumptions have been made in order to reduce the complexity of the situation.

The main difference is that the speech coming in for processing is not necessarily completely noiseless. Some distortions from the channel can be assumed, and in the worst case scenario, there is also environmental noise at the transmitting side of the mobile phone connection. The post-processing problem becomes far more difficult when the processed signal is already noisy because the noise is easily enhanced at the same time with the speech. Also, the estimation of formants is more demanding especially if the noise is not stationary. If the speech is affected by environmental noise in both ends of the communication channel, some kind of noise suppression would be needed before the post-processing block.

In the current realization, the post-filter needs information about the current frame as well as the next frame which means that 40 milliseconds of speech has to be buffered before the processing can be completed. The most time-consuming part of the post-processing is the interpolation of the filter coefficients which is done every 20th sample. To speed up the processing, the interpolation could probably be changed to every 40th sample without affecting the audible quality of the processed speech much. The problem is that the smoothing period between unvoiced and voiced frames is only 5 milliseconds long. This means that the whole subframe taken from the unvoiced frame would be filtered with a neutral filter, and then, at the beginning of the next voiced frame, the filter would suddenly change to the more drastic version. The memory of the post-filter would still be initialized with more

reasonable values than mere zeros, but the sudden change between frames is likely to cause some artifacts.

Besides the interpolation, the post-processing algorithm requires the following operations. For the pre-emphasis, a first order LP analysis is needed, and the pre-emphasis is done with a first order FIR filter. In an ideal situation, the decoder block would be able to pass the linear prediction coefficients of the speech frame to the post-filter. If this is not the case, a 10th order LP has to be calculated. The LP spectrum of the speech frame is formed by using a 256-sample FFT. After the formant filter has been formed, a first order LP is needed to determine the tilt. The final post-filter is a 5th order IIR filter. Of course, in addition to these, there are several steps, such as peak picking from the LP spectrum, which require some computations.

6.3 Further research

This section contains some ideas that could be further studied and also some possible improvements to the post-processing algorithm. Most of these ideas were invented during the writing of this thesis and their benefits remain unclear until tested. Also some other changes and features were overlooked during this phase of the work because the first goal was to merely get a working processing scheme that would provide positive results in terms of improved intelligibility at least in some situations.

In the current post-processing scheme, the filter parameters, r_i, are constant. Both of the numerator parameters were chosen to be 0.9 because this offers a good dynamic range for the filter. This does not necessarily mean that they are optimal, but it was too difficult to use subjective tests to optimize all four parameters. Initially, some objective measures, such as the Dau measure mentioned in Chapter 2, were considered for this purpose but the results were discouraging. The problem is that they do not reflect subjective preferences very well. Perhaps a combination of different measures or a more carefully defined optimization criterion could be utilized to achieve more beneficial results.

Even though the parameters r_i are constant, the filter constantly changes according to the formant frequencies. This also changes the filter gains, and formants in different frequency locations are enhanced or attenuated differently. In Chapter 5, it was concluded that the differences in gains are not very large, but it does open up another possible approach. Instead of defining the filter through the parameters r_i, it could have been defined through decibel gains on the first and second formant. The problem with this approach is that the dependence between the two is a rather complex mathematical equation. The calculation of the gains given the values of r_i is straightforward, but the other way around requires more computation. If the dependence could be simplified with an approximation that had a

relatively small error, the filter could be made more adaptive and intuitively more clear as decibel gains are much easier to understand than some arbitrary filter coefficients.

A few steps in the post-processing scheme were realized with rather simple and computationally inexpensive methods. They were deemed good enough since the problems in question, such as locating formant frequencies and separating between voiced and unvoiced speech frames, are extremely difficult. These parts could be further developed, not necessarily towards a much more complex realization, but a more accurate one. Of course, some complexity has to be added in order to improve the algorithms. Also, as mentioned earlier in Chapter 5, the post-filter actually enhances something near the resolved formants, and it would be beneficial to calculate the amount of this drift. It could be taken into account in the post-filtering, thus maybe further improving the results.

The next logical step would be conducting a formal subjective test to obtain some real measure for the performance of the post-filter. However, as was discussed in Chapter 2, the difficulty is in deciding whether the focus should be on quality or intelligibility. One simple solution is to conduct one of each. In a quality test, even a slightly negative result can be a good thing, if it is accompanied by a positive result on intelligibility. The ideal situation would be a scheme that improves intelligibility while maintaining or even improving the quality as well. Also the question of language remains. It is doubtful that the performance of the post-filter would be affected by the change of language from Finnish to English very much, but it would be interesting to see how the speech intelligibility index values discussed in Chapter 4 would be affected by the switch.

The subjective listening test that was conducted during this work cannot really be used to draw any kind of further conclusions. This means that since the test was designed solely for the purpose of optimizing parameter values, it is of little use elsewhere. The data is very scattered as was predicted, and it is hard to spot correlations between characteristics of the speakers and the corresponding parameter values. For this purpose, much more test subjects and speakers or a much more structured test would be needed. For example, two samples processed with different attenuations of the first formant would be given, and the listeners would be asked to pick their favorite. Once again, there would be a risk of getting random responses, if the differences between the samples were small and therefore inaudible to some listeners. But if succesful, this kind of test would produce more structured data which could be used to test whether the fundamental frequency of the speaker affects listener preferences and so on. As was discussed in Chapter 4, the correlations between the parameter values from the current test data and the F0 frequencies of the speakers are statistically insignificant. Other things that would be interesting to test include the effects of the first and second formant frequencies on the perceived quality and whether the filtering should be made adaptive to one or both of them.

The feedback loop from the noisy speech signal has not been realized yet. The system could be made adaptive to the level of environmental noise or even to the type of noise. In other words, when the SNR of the resulting noisy speech signal would be very low, the post-processing could be more extreme and then the effects of the post-processing would gradually decrease as the SNR increases. Or the filter could be turned on only after the signal-to-noise ratio has decreased past some limit. This way, the quality of the speech would not be affected if the conditions were good.

Also the noise type or the characteristics of the noise could have an effect on the post-processing. For now, the scheme has only been tested with car noise and briefly with office noise, but it can be assumed that it works well with stationary low-pass type noises. However, there might be some small differences in the optimal setting for separate noise types. The most difficult problem would probably be adopting the system to work well with babble noise. This noise type consists of multiple talkers speaking concurrently, and it is difficult to separate from the desired speech signal. With this type of noise the answer may not be in attenuating the first formant, but it is a good starting place. Hall et al. have already proposed a similar approach with promising results.

As mentioned earlier in this chapter, it has been constantly assumed that the speech signal that reaches the post-processing block is relatively noiseless. If this is not the case, the problem changes almost completely. Even though, the speech can be degraded to the point where the extraction of any information is extremely difficult and almost nothing can be done, the post-processing algorithm should take this kind of situation somehow into account. If the received speech signal has a very high noise level, then maybe the post-filtering should be turned completely off in order to avoid further enhancing the noise. If the situation was not as bad, some kind of noise suppression could be utilized as suggested earlier. In any case, this should also be further investigated, since in a real situation a completely noiseless speech signal is an unlikely occurrence.

Bibliography

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[2] I. V. McLoughlin, "Line spectral pairs," *Signal Processing*, vol. 88, no. 3, pp. 448–467, 2008.

[3] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 135–141, Apr 1974.

[4] I. Thomas and R. J. Niederjohn, "The intelligibility of filtered-clipped speech in noise," *Journal of the Audio Engineering Society*, vol. 18, no. 3, pp. 299–302, Jun 1970.

[5] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 277–282, Aug 1976.

[6] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Communication*, vol. 24, no. 3, pp. 211–226, 1998.

[7] M. Skowronski and J. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[8] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 280–285, 2010.

[9] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 59–71, Jan 1995.

[10] 3rd Generation Partnership Project (3GPP), "Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions, 3GPP TS 26.090," 2008, version 8.0.0.

[11] A. Mustapha and S. Yeldener, "An adaptive post-filtering technique based on the modified Yule-Walker filter," in *ICASSP '99: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999*. Washington, DC, USA: IEEE Computer Society, 1999, pp. 197–200.

[12] H. Kim and H.-G. Kang, "An adaptive short-term postfilter based on pseudo-cepstral representation of line spectral frequencies," *Speech Communication*, vol. 37, no. 3-4, pp. 335–348, 2002.

[13] H. Farsi, "A novel postfiltering technique using adaptive spectral decomposition for quality enhancement of coded speech," *WSEAS Transactions on Signal Processing*, vol. 4, no. 5, pp. 289–299, 2008.

[14] ITU-T, "Recommendation G.729 : Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic Code-Excited Linear Prediction (CS-ACELP)," March 1996.

[15] V. Grancharov, J. Plasberg, J. Samuelsson, and W. Kleijn, "Generalized postfilter for speech quality enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 57–64, Jan. 2008.

[16] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. model structure," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[17] W. Chen, P. Kabal, and T. Shabestary, "Perceptual postfilter estimation for low bit rate speech coders using Gaussian mixture models," in *INTERSPEECH-2005*, 2005, pp. 3161–3164.

[18] W. Chen, "Perceptual postfiltering for low bit rate speech coders," Master's thesis, McGill University, 2007.

[19] ITU-T, "Recommendation G.723.1 : Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," May 2006.

[20] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[21] ITU-T, "Recommendation P.862 : Perceptual evaluation of speech quality (PESQ) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," February 2001.

[22] ITU-T, "Recommendation P.800 : Methods for subjective determination of transmission quality," August 1996.

[23] ITU-R, "Recommendation BS.1534-1 : Method for the subjective assessment of intermediate quality level of coding systems," January 2003.

[24] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *The Journal of the Acoustical Society of America*, vol. 30, no. 7, pp. 596–600, 1958.

[25] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[26] ANSI, "Methods for calculation of the speech intelligibility index (ANSI S3.5)," June 1997.

[27] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[28] B. Sauert, G. Enzner, and P. Vary, "Near End Listening Enhancement with Strict Loudspeaker Output Power Constraining," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2006.

[29] ITU-T, "Recommendation G.191 : Software tools for speech and audio coding standardization," September 2005.

[30] S. F. de Campos Neto, "The ITU-T software tool library," *International Journal of Speech Technology*, vol. 2, pp. 259–272, 1999.

[31] ITU-T, "Recommendation P.56 : Objective measurement of active speech level," March 1993.

[32] ITU-T Users' Group on Software Tools, *ITU-T Software Tool Library 2005 User's Manual*, August 2005.

[33] 3rd Generation Partnership Project (3GPP), "ANSI-C code for the floating-point adaptive multi-rate (AMR) speech codec, 3GPP TS 26.104," 2009, version 9.0.0.

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

[35] L. Laaksonen, J. Kontio, and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, Mar. 2005, pp. 809–812.

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

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

[38] Acoustical Society of America Working Group S3-79. SII : Speech intelligibility index. Retrieved : 10.07.2009. [Online]. Available: http://sii.to/

[39] A. Härmä and K. Palomäki. (1999, October) HUTear - Matlab toolbox for auditory modeling. Retrieved : 08.06.2009. [Online]. Available: http://www.acoustics.hut.fi/software/HUTear/

[40] C. H. Shadle, "Experiments on the acoustics of whistling," *The Physics Teacher*, vol. 21, no. 3, pp. 148–154, March 1983.

[41] K. Wu and D. G. Childers, "Gender recognition from speech. Part I: Coarse analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[42] L. A. Thorpe, "Subjective evaluation of speech compression codecs and other nonlinear voice-path devices for telephony applications," *International Journal of Speech Technology*, vol. 2, pp. 273–288, May 1999.

[43] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1-2, pp. 1–194, 2007.

[44] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing - A Practical Approach*. Addison-Wesley Longman Publishing Co., 1993.

[45] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice-Hall, Inc., 1996.

[46] S. K. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill Higher Education, 2000.

[47] P. Alku, "Puheensiirtotekniikka," January 2009, lecture notes.

Appendix A

Derivation of the filter amplitude response

In the following the amplitude response of the post-filter structure is derived step by step.

$$
\begin{aligned}
|H(z)|^2 &= \left| \frac{1 - 2r_{11}cos\theta_1 z^{-1} + r_{11}^2 z^{-2}}{1 - 2r_{12}cos\theta_1 z^{-1} + r_{12}^2 z^{-2}} \cdot \frac{1 - 2r_{21}cos\theta_2 z^{-1} + r_{21}^2 z^{-2}}{1 - 2r_{22}cos\theta_2 z^{-1} + r_{22}^2 z^{-2}} \cdot \frac{1}{1 + \mu z^{-1}} \right|^2 \\
&= \left| \frac{1 - 2r_{11}cos\theta_1 z^{-1} + r_{11}^2 z^{-2}}{1 - 2r_{12}cos\theta_1 z^{-1} + r_{12}^2 z^{-2}} \right|^2 \cdot \left| \frac{1 - 2r_{21}cos\theta_2 z^{-1} + r_{21}^2 z^{-2}}{1 - 2r_{22}cos\theta_2 z^{-1} + r_{22}^2 z^{-2}} \right|^2 \cdot \left| \frac{1}{1 + \mu z^{-1}} \right|^2
\end{aligned}
$$

Thus, the filters can be treated separately. Both of the formant filters are of the same form:

$$
H(z) = \frac{1 - 2r_1 cos\theta z^{-1} + r_1^2 z^{-2}}{1 - 2r_2 cos\theta z^{-1} + r_2^2 z^{-2}}
$$

When $z = e^{j\omega}$,

$$
\begin{aligned}
H(e^{j\omega}) &= \frac{1 - 2r_1 cos(\theta)cos(\omega) + j2r_1 cos(\theta)sin(\omega) + r_1^2 cos(2\omega) - jr_1^2 sin(2\omega)}{1 - 2r_2 cos(\theta)cos(\omega) + j2r_2 cos(\theta)sin(\omega) + r_2^2 cos(2\omega) - jr_2^2 sin(2\omega)} \\
&= \frac{1 - 2r_1 cos(\theta)cos(\omega) + r_1^2 cos(2\omega) + j(2r_1 cos(\theta)sin(\omega) - r_1^2 sin(2\omega))}{1 - 2r_2 cos(\theta)cos(\omega) + r_2^2 cos(2\omega) + j(2r_2 cos(\theta)sin(\omega) - r_2^2 sin(2\omega))} \\
&= \frac{A + jB}{C + jD} \\
&= \frac{(A + jB) \cdot (C - jD)}{(C + jD) \cdot (C - jD)} \\
&= \frac{AC + BD + j(BC - AD)}{C^2 + D^2} \\
|H(e^{j\omega})|^2 &= \left| \frac{AC + BD + j(BC - AD)}{C^2 + D^2} \right|^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(AC+BD)^2+(BC-AD)^2}{(C^2+D^2)^2} \\
&= \frac{A^2C^2+2ABCD+B^2D^2+B^2C^2-2ABCD+A^2D^2}{(C^2+D^2)^2} \\
&= \frac{(A^2+B^2)\cdot(C^2+D^2)}{(C^2+D^2)^2} \\
&= \frac{A^2+B^2}{C^2+D^2} \\
&= \frac{(1-2r_1cos(\theta)cos(\omega)+r_1^2cos(2\omega))^2+(2r_1cos(\theta)sin(\omega)-r_1^2sin(2\omega))^2}{(1-2r_2cos(\theta)cos(\omega)+r_2^2cos(2\omega))^2+(2r_2cos(\theta)sin(\omega)-r_2^2sin(2\omega))^2} \\
&= \frac{1+r_1^4+4r_1^2cos^2(\theta)-4r_1cos(\theta)cos(\omega)(r_1^2+1)+2r_1^2cos(2\omega)}{1+r_2^4+4r_2^2cos^2(\theta)-4r_2cos(\theta)cos(\omega)(r_2^2+1)+2r_2^2cos(2\omega)}
\end{aligned}
$$

The tilt compensation filter has the form:

$$\Gamma(z) = \frac{1}{1-\mu z^{-1}}$$

When $z = e^{j\omega}$,

$$
\begin{aligned}
\Gamma(e^{j\omega}) &= \frac{1}{1-\mu cos(\omega)+j\mu sin(\omega)} \\
&= \frac{1-\mu cos(\omega)-j\mu sin(\omega)}{(1-\mu cos(\omega))^2+(\mu sin(\omega))^2} \\
\left|\Gamma(e^{j\omega}\right|^2 &= \left|\frac{1-\mu cos(\omega)-j\mu sin(\omega)}{(1-\mu cos(\omega))^2+(\mu sin(\omega))^2}\right|^2 \\
&= \frac{(1-\mu cos(\omega))^2+(\mu sin(\omega))^2}{((1-\mu cos(\omega))^2+(\mu sin(\omega))^2)^2} \\
&= \frac{1}{1-2\mu cos(\omega)+\mu^2(cos^2(\omega)+sin^2(\omega))} \\
&= \frac{1}{1-2\mu cos(\omega)+\mu^2}
\end{aligned}
$$

Appendix B

Test sentences

Table B.1 presents the sentences that were used in the subjective listening test. They were extracted from a longer Finnish text.

#	Speaker	Gender	Sentence
1	HaPu	M	Vuodenvaihdetta juhlitaan talvisäässä.
2	MaAi	M	Heinäkuun viimeinen päivä oli äärimmäisen harvinainen.
3	PaAl	M	Jääkartat ovat nyt päivittäin saatavilla laitoksen verkkosivuilla.
4	HeLe	F	Sademäärä oli noin 150 millimetriä.
5	LaLe	F	Kuukauden keskilämpötilat lähestyivät lämpöennätyksiä.
6	VeAl	F	Tutkimuksen avulla saadaan päätöksentekijöille tietoa ilmanlaadusta.

Table B.1 – The Finnish sentences used in the listening test.

Appendix C

Test instructions

Below are the instructions that were given to the listeners in the subjective listening test. The instructions are in Finnish.

Kuuntelukoe

Tässä kokeessa kuuntelet lyhyitä lauseita. Koe jakautuu kuuden näytteen kuunteluun, joista jokaisen kohdalla tehtävänäsi on valita paras käsittely kyseiselle näytteelle.

Kokeessa käytettävä käyttöliittymä on kuvan C.1 mukainen. Valkoisen alueen reunassa on vaaleansininen ”NEUTRAALI”-nappi, jota klikkaamalla voi kuunnella alkuperäisen äänen. Edettäessä akseleita pitkin kauemmaksi neutraalista näytteestä äänelle tehtävä käsittely voimistuu. Klikkaamalla valkoista aluetta saadaan kuultavaksi ääni, jota on käsitelty sen sijaintia vastaavalla tavalla, ja klikattuun kohtaan ilmestyy punainen neliö. Ruudulle ilmestyneitä neliöitä klikkaamalla voi kuunnella uudelleen vastaavassa kohdassa sijaitsevan äänen. Punainen neliö kertoo aina viimeksi kuunnellun äänen sijainnin. **Tarkoituksena on hakea valkoiselta alueelta kohta, joka tuottaa mielestäsi mahdollisimman selkeän, ymmärrettävän ja miellyttävän version äänestä.**

”SEURAAVA”-nappia painamalla voi siirtyä eteenpäin seuraavaan näytteeseen. Tällöin punaisella neliöllä merkitty ääni tallentuu kuuntelijan valinnaksi ja seuraavan näytteen arviointi alkaa. Aiempiin näytteisiin palaaminen ei ole mahdollista, joten varmista, että valintasi mukainen ääni on merkitty punaiseksi, ennen kuin etenet seuraavaan näytteeseen.

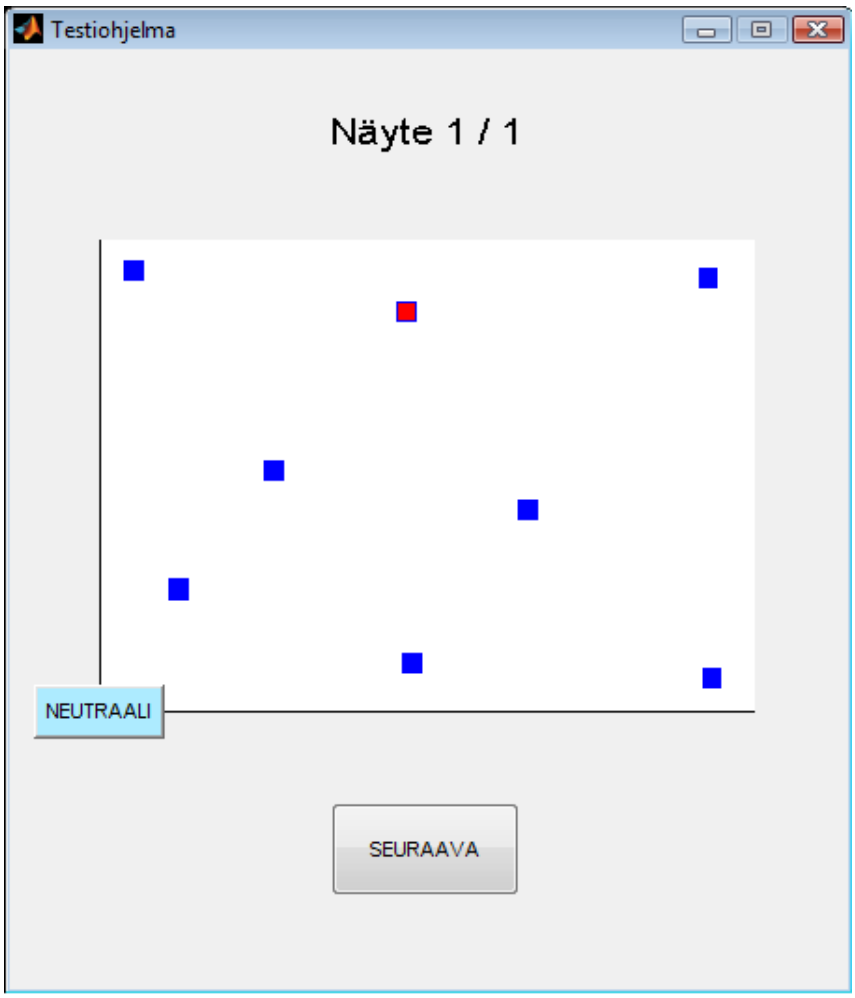


Figure C.1 – Käyttöliittymä.

Appendix D

Example of data from the subjective tests

Figure D.1 shows the clicks made by one of the test subjects in the subjective listening test. The order of the samples is the same as their playing order for that particular subject. The red marker is the final choice for that sample.

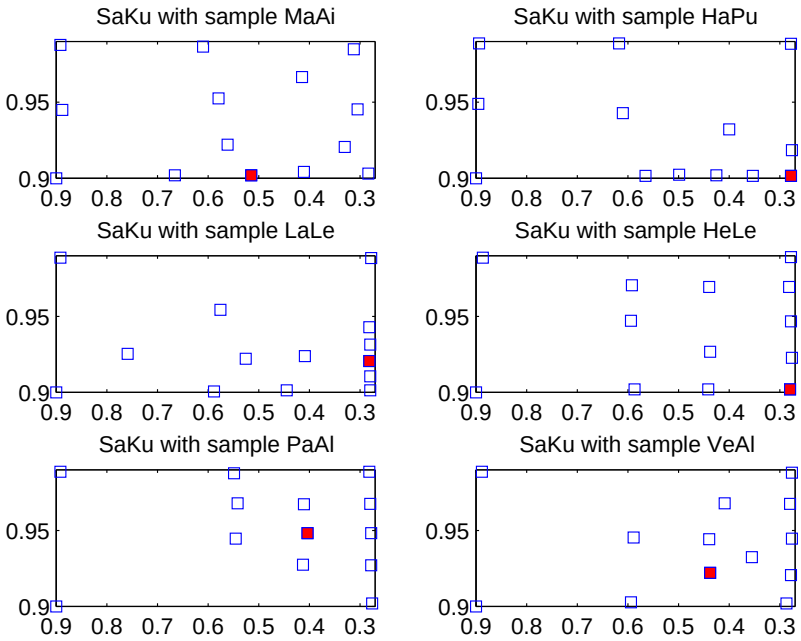


Figure D.1 – Data obtained from the listening tests with subject SaKu.