

Decoder based noise suppression

Erik Hennix

January 19, 2006

Abstract

Acoustic background noise in mobile speech communication systems, while largely inevitable, can have a severely detrimental effect on speech intelligibility. Noise suppression is highly desirable in these systems. However, the process of reducing noise in a speech signal is associated with distortion of the processed signal, the severity of which is generally proportional to the amount of noise suppression applied. Placing the noise suppression at the decoder side of a communication link, allows for user control, but for a speech coder to be used for coding noisy signals it must be robust to noise. The thesis explores these two issues regarding this type of system: 1) To evaluate the feasibility of decoder based noise suppression, the effect of coding noisy speech signals with the Adaptive Multi Rate codec (AMR) at several bit-rates is evaluated for various noise types and noise levels, using objective methods. The results show that the coding is robust to noise and has a signal enhancing effect in these tests. 2) User control is simulated using the noise suppression system present in the Enhanced Variable Rate Codec (EVRC) and user control over several system parameters is implemented and evaluated using subjective tests. The results of these tests show significant differences in individual preferences.

Acknowledgements

This thesis was made possible by the help and contributions of many people. I would first like to thank my supervisor Volodya Grancharov for giving me the opportunity to do this thesis and his invaluable support and guidance throughout the course of my work. I also want to thank Jonas Samuelsson for suggesting this thesis work to me and giving me valuable feedback on my report. I would like to express my deep gratitude to the kind people at S3 who took the time to participate as subjects in the listening test for a meager reward. Last, but not least, I would like to thank my friends, my family and Cecilia for their continuing love, support and encouragement.

Contents

1	Introduction	2
1.1	The need for noise suppression	2
1.2	Noise suppression system configurations	3
1.3	Motivation and objectives	4
1.4	Outline of contents	5
2	Noise suppression	6
2.1	Classifying noise suppression methods	6
2.2	Spectral subtraction	7
3	Noise estimation	10
3.1	Voice activity detection	10
3.2	Quantile based estimation	11
4	Speech coding	12
4.1	The source-filter model of speech production	12
4.2	Code excited linear prediction	13
4.3	Vector quantization of spectral coefficients	14
4.4	Spectral effects of coding noisy voiced speech	15
5	The EVRC noise suppression system	18
5.1	Initial time domain processing	19
5.1.1	Pre-emphasis	19
5.1.2	Windowing	19
5.1.3	Frequency domain conversion	20
5.2	Frequency domain processing	20
5.2.1	Channel energy estimator	22
5.2.2	Channel SNR estimator	22
5.2.3	Voice metric calculation	22
5.2.4	Spectral deviation and power spectral estimator	23
5.2.5	SNR estimate modifier	23

5.2.6	Channel gain calculation	23
5.2.7	Frequency domain filtering	24
5.2.8	Background noise estimate update	25
5.3	Final time domain processing	25
5.3.1	Time domain conversion	25
5.3.2	Overlap-add and de-emphasis	26
6	User controlled noise suppression	27
6.1	Controlling the EVRC noise suppression system	27
7	Evaluation	29
7.1	The effect of coding a noisy speech signal	29
7.1.1	Post coder noise suppression	31
7.2	User controlled noise suppression	32
7.3	Results	34
8	Conclusions and suggested future work	46
8.1	Conclusions	46
8.2	Suggested future work	47
A	EVRC Noise Suppression	48
A.1	The channel combining tables	48
A.2	The voice metric table	48
A.3	The exponential windowing factor, $\alpha(m)$	48
A.4	SNR Estimate modification	49
A.5	Noise update decision	50

List of Figures

1.1	A speech communication system with a pre-processor noise suppression system (Conventional configuration).	3
1.2	A speech communication system with a post-processor (decoder based) noise suppression system.	4
4.1	The general principle of vector quantization.	15
4.2	LP spectrum of voiced segment A.	16
4.3	LP spectrum of voiced segment B.	17
4.4	LP spectrum of voiced segment C.	17
5.1	General principle of the EVRC NS system.	18
5.2	EVRC NS initial time domain processing block.	19
5.3	EVRC-NS input buffer overlapping and pre-emphasis.	20
5.4	EVRC NS frequency domain processing block.	21
5.5	Single channel gains for EVRC-NS and Spectral subtraction.	24
5.6	EVRC NS final time domain processing block.	26
7.1	Evaluating a system with PESQ using clean speech input.	30
7.2	Evaluating a system with PESQ using noisy speech input.	30
7.3	The setup used for evaluating the signal quality of the noisy speech signals.	31
7.4	The setup used for evaluating the signal quality of the processed (coded-decoded) noisy speech signals.	31
7.5	The setup used for evaluating post coder noise suppression.	32
7.6	The setup of the user control listening test.	32
7.7	The graphical user interface of the test software used to evaluate user control over EVRC noise suppression.	33
7.8	PESQ results for white, factory and babble noise.	35
7.9	PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. White noise added.	36

7.10	PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. Factory noise added.	37
7.11	PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. Babble noise added.	37
7.12	Listener preferences for the channel energy smoothing factor, α_{ch} , female speaker.	38
7.13	Listener preferences for the channel energy smoothing factor, α_{ch} , male speaker.	38
7.14	Listener preferences for the noise smoothing factor, α_n , female speaker.	39
7.15	Listener preferences for the noise smoothing factor, α_n , male speaker.	39
7.16	Listener preferences for the minimum overall gain, γ_{min} , female speaker.	40
7.17	Listener preferences for the minimum overall gain, γ_{min} , male speaker.	40
7.18	Preferences for the channel energy smoothing factor, α_{ch} , for all subjects.	43
7.19	Preferences for the noise smoothing factor, α_n , for all subjects.	44
7.20	Preferences for the minimum overall gain, γ_{min} , for all subjects.	45

List of Tables

7.1	The opinion scale recommended by ITU-T for listening quality tests.	29
7.2	PESQMOS scores for noisy speech and coded-decoded noisy speech. White noise added.	35
7.3	PESQMOS scores for noisy speech and coded-decoded noisy speech. Factory noise added.	35
7.4	PESQMOS scores for noisy speech and coded-decoded noisy speech. Babble noise added.	36
7.5	Test subjects' preferences for the channel energy smoothing factor, α_{ch}	41
7.6	Test subjects' preferences for the minimum overall gain, γ_{min}	42
7.7	Test subjects' preferences for the noise smoothing factor, α_n	42

Chapter 1

Introduction

1.1 The need for noise suppression

When communicating via telephone it is inevitable that the microphone in the transmitting terminal picks up, at least to some extent, any acoustic background noise present, in addition to the speech sound. In civilian applications, this has tended to be fairly low level noise. Severe additive acoustic noise used to be the concern of engineers of military communication systems. Indeed, many early papers detailing noise suppression techniques are aimed at reducing noise of the type encountered in helicopter- and jet aircraft cockpits [2][3].

In more recent times the widespread use of digital mobile phones has brought these problems into the focus of civilian telecommunications engineering. Many cell-phone conversations are carried out in cars, on public transportation and on busy city streets, with any number of acoustic noise sources present at levels that can be quite high. Although human perception of speech is highly resistant to additive noise [8], it is annoying and can reduce speech intelligibility. Thus, there is a need for noise suppression in mobile communication systems.

An area of digital speech signal processing where noise suppression is a key component, is automatic speech recognition and speaker authentication. The error rate of these systems increases dramatically with the presence of noise in the input signal [9]. This is unfortunate, since environments where these systems can be assumed to be helpful include noisy environments, such as vehicles and factories for speech recognition, or streets and driveways for speaker authentication. Effective noise suppression that can handle a wide range of non stationary noise is highly desirable as a pre-processor to these types of systems.

1.2 Noise suppression system configurations

The noise suppression (NS) systems used in telecommunications today are mostly of the type designed to operate on time domain (TD) signal samples. As such these systems are fairly stand alone processing blocks. In the conventional configuration (see figure 1.1), assuming a system with noise suppression, NS processing is performed on the signal before coding it for transmission. This has the advantage of presenting the coder with a less noisy signal, which can improve its performance. However, if the coder is robust to noisy input, the NS system could be placed at the decoder (receiver) side, as in figure 1.2. This has the advantage of enabling user control over the noise suppression system. The terminal at the receiving side could be equipped with a control for noise suppression, much like the volume control of mobile handsets today. Also, as is discussed in section 4.4, the speech coding process can have a signal enhancing effect on noisy speech. A decoder based noise suppression system could be designed to exploit this.

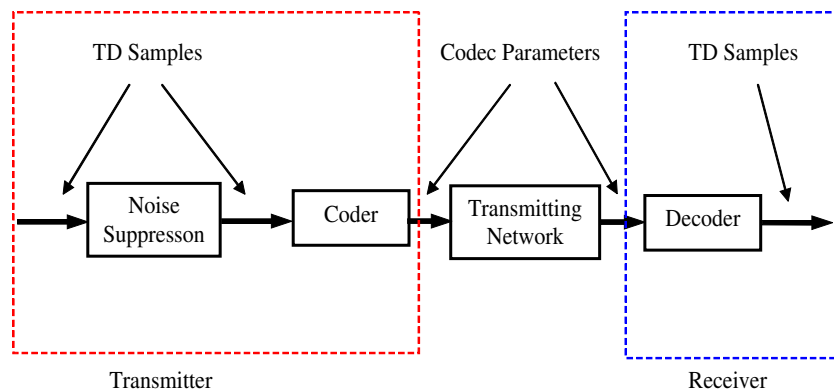


Figure 1.1: A speech communication system with a pre-processor noise suppression system (Conventional configuration).

Systems working on codec parameters have also been suggested [10][11]. These can be placed either in the transmitting network or before the decoder in the receiver. The implementation of NS on codec parameters differs in many ways from systems operating on signal samples, and is beyond the scope of this thesis.

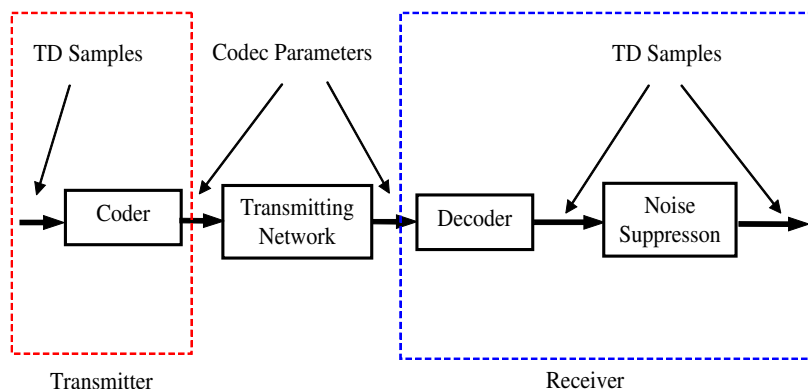


Figure 1.2: A speech communication system with a post-processor (decoder based) noise suppression system.

1.3 Motivation and objectives

Acoustic background noise can have a severely detrimental effect on speech intelligibility and perceived signal quality in mobile communication systems. Effective noise suppression is a very desirable feature for these and many other speech signal applications. This thesis will focus on noise suppression in a mobile phone context, using existing commercial systems.

Unfortunately, NS processing introduces distortion, the magnitude of which tends to increase with the magnitude of the noise suppression. User control would enable the listener to adjust the trade-off between distortion and noise level according to his/her preference. Placing the NS processing at the decoder (receiver) allows user control, but requires the coder to be able to handle a noisy input signal. Thus, speech coding will be studied and the effect of coding a noisy speech signal will be evaluated.

User controlled noise suppression will be evaluated in the context of a simulated application. For this, a state-of-the-art commercial NS system will be thoroughly studied. Using objective measures, it will be shown to be able to function in a post processor configuration. While the motivation and implementation for commonplace user sound controls such as volume and tone (bass-treble) are relatively straight-forward, the case for noise suppression is not. Taking as starting point our chosen NS system, a small number of parameters of the system will be selected for testing user control, based on their subjective impact. The evaluation will be in the form of a series of listening tests.

1.4 Outline of contents

In Chapter 2, an overview of noise suppression techniques and their classification is given. Special attention is given to Spectral subtraction, which is the basis for many current noise suppression systems. The next chapter focuses on the problem of noise estimation, which is a difficult but essential part of any noise suppression system. In Chapter 4, speech coding and the effects of coding noisy speech is explored. Chapter 5 is a detailed description of the noise suppression system present in the Enhanced Variable Rate Codec and the following chapter discusses user control over this particular NS system. The evaluation of coding noisy speech is presented in Chapter 7, as well as the details and results of a listening test that simulated user control of the EVRC NS system. Conclusions and suggested future work are given in Chapter 8. The appendix contains details on specific parts of the EVRC NS system.

Chapter 2

Noise suppression

2.1 Classifying noise suppression methods

Today, there exists a multitude of noise suppression methods for speech signals, and the ways to classify them are many as well. They can be either time domain methods, such as [6] [7], which work directly on the signal samples, or frequency domain methods, in which the signal is first transformed, via e.g. the Fourier transform. All processing is then carried out on the frequency domain representation of the signal, and the result is finally transformed back to the time domain.

The noise suppression systems described in this thesis are based on the noise suppression system present in the Enhanced Variable Rate Codec (EVRC) [15], which uses a frequency domain method to suppress background noise. A detailed description of the EVRC noise suppression system can be found in section 5, while the general theory behind the method (spectral subtraction) is discussed later in this chapter.

Any signal processing scheme is ultimately limited by the available input signal(s). Multi channel systems, with several microphones, can use beamforming techniques to achieve a higher signal to noise ratio by directing the pick up at the speaker. With a multi channel system using two microphones, one microphone can be placed close to the source of the desirable signal, while the other is placed a bit further away, so that it mostly picks up background noise. The noise in the microphone close to the source, called the primary signal, is correlated to that from the other microphone, called the reference signal. In adaptive noise cancellation (ANC), the reference is filtered, using an adaptive filter, and subtracted from the primary signal [16]. The systems described in this thesis are all single channel applications, multichannel techniques are not used.

One can also separate systems into parametric and non-parametric systems. Parametric systems, e.g. [4] [5], map the input signal onto a parameter space according to some model. The processing is performed on the parameters and the result is used to reconstruct the output signal. Non-parametric systems work directly on the time domain signal samples or some transformation of the signal that is not model based, e.g. the discrete fourier transform (DFT) of short blocks of the signal. The noise suppression method used in this thesis is non-parametric.

2.2 Spectral subtraction

Spectral subtraction is the method used in the noise suppression systems in this thesis. It is a frequency domain, non-parametric method which attempts to reduce the average effects of additive acoustic noise by subtracting an estimate of the noise spectrum from the input signal. One of the defining papers describing spectral subtraction as a means for suppressing acoustic noise in a speech signal was written by Steven Boll in 1979 [2]. Though much research has been made in the field of speech enhancement and noise suppression, the method described by Boll is still the basis for many contemporary noise suppression systems. This is likely due to the fact that spectral subtraction is a fairly robust, low complexity method. However, although it is straightforward to implement in principle, it is far from a perfect method and care must be taken to minimize distortion and artifacts in the processed signal.

Consider the single channel noise reduction problem, where the desired signal $s(n)$ has been corrupted with additive noise, $v(n)$, which is assumed to be uncorrelated with $s(n)$, according to

$$y(n) = s(n) + v(n). \quad (2.1)$$

The input signal, $y(n)$, is processed in short segments, frames, which are generally 10-30 ms in duration [13]. It is assumed that for the duration of a frame, $s(n)$ and $v(n)$ can be considered to be wide sense stationary. Spectral subtraction can be performed on either the short time magnitude spectrum of $y(n)$, $|Y(\omega)|$, or the power spectrum $|Y(\omega)|^2$. The idea is to subtract an estimate of the short time magnitude or power spectrum of the noise from that of the input signal, to obtain an estimate of the clean signal spectrum. For power spectral subtraction

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{V}(\omega)|^2; & \text{if } |Y(\omega)|^2 > |\hat{V}(\omega)|^2 \\ 0; & \text{otherwise} \end{cases}, \quad (2.2)$$

where $|\hat{S}(\omega)|^2$ is the estimated short-time power spectrum of the clean signal and $|\hat{V}(\omega)|^2$ is the time-averaged estimate of the short-time power spectrum of the noise. Power spectral subtraction can also be expressed, and is commonly implemented as, a multiplication with a time varying filter according to

$$|\hat{S}(\omega)|^2 = H(\omega) \cdot |Y(\omega)|^2 \quad (2.3)$$

where

$$H(\omega) = \begin{cases} 1 - \frac{|\hat{V}(\omega)|^2}{|Y(\omega)|^2}; & \text{if } |Y(\omega)|^2 > |\hat{V}(\omega)|^2 \\ 0; & \text{otherwise} \end{cases} . \quad (2.4)$$

Spectral components of the input signal that have a high level compared to the noise are emphasized, as components are more attenuated the closer their levels are to that of the noise. The phase of the input signal is not modified, thus the estimate of the clean signal will still have the phase of the noisy signal. This has little effect on the perceived quality of the resulting signal, since human hearing is relatively insensitive to phase [8]. Other factors have a far greater impact. Firstly, it is critical for the noise estimate to be accurate. Noise estimation is a major problem in itself and is the subject of the next chapter. Suffice to say that a deviation in the noise estimate from the real noise will result in some residual noise in the output signal. One of the major drawbacks with subtractive noise suppression is that the residual noise tends to be modulated. It can have a tone-like quality which is annoying to a human listener. This is *musical* residual noise. Several techniques are used to minimize the presence of musical residual noise. One is to apply over subtraction, which can be expressed as multiplying the SNR estimate with a factor $\alpha > 1$ when calculating the subtraction filter, so that the filter takes the form

$$H_{OS}(\omega) = \begin{cases} 1 - \alpha \cdot \frac{|\hat{V}(\omega)|^2}{|Y(\omega)|^2}; & \text{if } |Y(\omega)|^2 > |\hat{V}(\omega)|^2 \\ 0; & \text{otherwise} \end{cases} . \quad (2.5)$$

This will attenuate the short-term spectrum more, causing a reduction of the residual noise peaks [17], but increasing the distortion of the signal. Another method, which does not exclude the use of over subtraction, is spectral flooring, whereby the subtraction filter is limited to a minimum value $H_{min} > 0$. This results in more background noise being present in the signal, which can have a masking effect on the residual noise.

Other ways to formulate the noise suppression rule detailed above have also been suggested, e.g. [18]. These have been shown to be able to reduce some of the artifacts associated with spectral subtraction, but at the cost of added complexity compared to the basic spectral subtraction method.

Although presented well over a decade ago, these methods have yet to be implemented widely in commercial speech communication systems. Indeed, the noise suppression system which is the focus of this thesis, i.e. the noise suppression present in EVRC, is based on a very simple suppression rule. Low complexity is a high priority in speech communication systems.

Chapter 3

Noise estimation

Thus far we have focused on how to reduce the noise in a signal once an estimate of the noise spectrum has been obtained. As mentioned, the method of spectral subtraction is highly dependant on the quality and validity of this estimate. Finding an estimate of the background noise is a central problem in the field of noise suppression.

3.1 Voice activity detection

The most common way to produce this estimate is based on the fact that speech has frequent pauses. During these the signal can be assumed to contain only noise and the noise spectrum estimate can be obtained as the time averaged spectrum of the input signal. To prevent rapid changes in the noise estimate, which could cause annoying artifacts in the processed signal, it is usually updated using a smoothing algorithm. Since the noise estimate can only be updated in speech pauses, it is likely to differ from the actual noise during speech segments, unless the background noise is completely stationary. Also, detecting speech pauses is not a trivial problem, especially at high noise levels. Nevertheless, techniques based on this method, known as voice activity detection (VAD), are widely used in noise suppression systems today, as they offer fairly good performance and low complexity. The noise spectrum estimation for the noise suppression system in EVRC is based on a VAD scheme (see section 5 for a detailed description). In more recent publications, e.g. [9], other methods have been presented, which allow the noise estimate to be continuously updated.

3.2 Quantile based estimation

In [9] Stahl et al present a method called quantile based noise estimation. This method is based on the statistical fact that in a speech signal, even during segments containing speech, the energy in any given frequency band is at the noise level during a significant percentage of the time. Consider a speech signal with power spectrum $|S(\omega_k, t)|^2$ that is contaminated with additive noise with power spectrum $|V(\omega_k)|^2$ to form the observed signal $|Y(\omega_k, t)|^2$ according to

$$|Y(\omega_k, t)|^2 = |S(\omega_k, t)|^2 + |V(\omega_k)|^2, \quad (3.1)$$

where ω_k represents a frequency band and t is the frame index. For each frequency band, the power spectra of the frames of the observed signal $|Y(\omega_k, t)|^2$, $t = 0..T$ are sorted such that

$$Y(\omega_k, t_0) \leq Y(\omega_k, t_1) \leq \dots \leq Y(\omega_k, t_T). \quad (3.2)$$

The noise power spectrum can now be estimated by taking the q^{th} quantile over time in each frequency band:

$$\hat{V}(\omega_k) = Y(\omega_k, t_{\lfloor qT \rfloor}), \quad (3.3)$$

where $q = 1$ yields the maximum, $q = 0$ the minimum and $q = 0.5$ the median. Stahl et al found that $q \approx 0.5$ seems to provide the best noise spectrum estimate. The advantages of being able to continuously update the noise and not having to detect speech pauses are offset by the added complexity and memory requirements inherent in this method. Also, the method described above is designed for stationary noise. If the noise is non-stationary, the noise estimation should be based on observations stored in a buffer of limited length. Choosing the length of such a buffer is a trade-off between how accurate the estimation will be for stationary noise and how quickly the system will adapt to changes in the noise.

Chapter 4

Speech coding

In order to transmit an audio signal over a digital channel, it has to be coded into bits. In its most trivial form, this coding can consist of sample amplitude values encoded as binary numbers of, e.g., 8-bits each. This raw data format, known as Pulse Code Modulation (PCM), is quite inefficient though. For efficient use of available network resources, the data rate (bits per second, bit-rate) should be as low as possible. More sophisticated coding strategies can lead to significant reductions in bit-rate, with minor loss of signal quality. When coding speech for telecommunication, it is not necessary for the coded (processed) signal to be completely true to the original, as long as it is clear, intelligible and sufficiently natural sounding. For transmission of music, e.g., the demands on fidelity are generally much higher. Coders for high-fidelity applications are outside the scope of this text.

4.1 The source-filter model of speech production

The speech coders used in this thesis work by finding a set of parameters that can be used to reconstruct the input speech waveform, according to a model of the human speech production system, rather than coding the actual sample values directly. The basic assumption is that short segments of speech can be modelled as the response of a linear quasi-stationary system to an excitation [12]. The linear system is the vocal tract, modelled as an all-pole filter of order p , with a transfer function of the form

$$V(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (4.1)$$

The excitation is the output of the vocal cords (larynx), which is periodic in voiced speech and noise-like in unvoiced speech. Thus, the speech signal is modelled as

$$s(n) = \sum_{k=1}^p a_k s(n-k) + u(n) + e(n) \quad (4.2)$$

where $u(n)$ represents the excitation from the vocal cords and $e(n)$ is the modelling error. To use this model in a speech coder, the parameters representing the vocal tract, a_k , must first be estimated. It is assumed that the vocal tract can be considered to be stationary for up to 20 ms at a time. Accordingly, the vocal tract parameters are estimated for signal frames (blocks of data) of about 10-20 ms in duration. Furthermore, a 10th order system is normally used to represent the vocal tract, i.e. $p = 10$. The excitation, $u(n)$, and the modelling error, $e(n)$, are combined in one single error term, $g \cdot w(n)$, where g is a gain factor and $w(n)$ is taken to be white Gaussian noise with zero mean and unit variance. The vocal tract parameters are estimated as the optimal parameters in the Minimum Mean Square Error (MMSE) sense. This criterion leads to a set of linear equations, for which there are several efficient solution methods [13].

4.2 Code excited linear prediction

The codecs used in this thesis include the Adaptive Multi Rate (AMR) [1] and the Enhanced Full Rate Codec (EFR) [14], which is used in the GSM digital cellular telephone system and is also the basis for AMR at the 12.2 kbits/s rate. They are all based on Code Excited Linear Prediction (CELP). At the coder (transmitter), for each input signal frame, the Linear Prediction Coefficients (LPCs) are computed. These are quantized and transmitted to the decoder (receiver). By filtering the input signal through the filter defined by the LPCs, the residual part of the signal is obtained. This signal, the *excitation*, needs to be transmitted for the decoder to be able to reconstruct the input speech. Code Excited means that the residual is quantized using Vector Quantization (VQ), so that the residual parameters are actually transmitted in the form of indices into a codebook. The quantization of the residual in EFR and AMR involves both an adaptive and a fixed codebook, and is outside the scope of this text. Relevant to this thesis, however, is that in these codecs the LPCs are also quantized using VQ. As well as allowing for efficient quantization, this can also have an enhancing effect on the spectral envelope of the input signal.

4.3 Vector quantization of spectral coefficients

The principle for vector quantization of spectral coefficients is shown in figure 4.1. The coder and decoder share a codebook of N vectors. At the coder, the quantizer compares the input vector, \mathbf{b} , with the vectors stored in the codebook, \mathbf{a} . The output of the quantizer is the index, i , of the codebook vector which is closest, according to some measure, to the input vector. This measure is called the distortion measure. Because LP coefficients are sensitive to quantization errors, they are not quantized directly. Instead, the LP coefficients are first transformed to the corresponding Line Spectral Frequencies (LSF), and these are then used as input to the quantizer. The vector quantization in EFR is performed by finding the codebook index, i , which minimizes the distortion measure, E_{LSF} , according to

$$E_{LSF} = \sum_{k=1}^{10} w_k (f_k - \hat{f}_k^i)^2, \quad (4.3)$$

where f_k is the k^{th} input LSF, \hat{f}_k^i is the k^{th} quantized LSF at index i and w_k is the k^{th} weighting factor. The weighting factors are given by

$$w_k = \begin{cases} 3.347 - \frac{1.547}{450} d_k & \text{for } d_k < 450 \\ 1.8 - \frac{0.8}{1050} (d_k - 450) & \text{otherwise} \end{cases} \quad (4.4)$$

where $d_k = f_{k+1} - f_{k-1}$ with $f_0 = 0$ and $f_{11} = 4000$. The codec exploits the fact that the LSF representations of the spectral coefficients are grouped in pairs close to the spectral peaks (formants), and weighs the errors at the peaks more than those in the valleys. This improves the performance of the spectral VQ for speech input and, as a side effect, it also makes the quantizer robust to noisy speech input.

At the decoder, the output is calculated as the LPC vector that corresponds to the i^{th} LSF vector in the codebook. Thus, the output is limited to the vectors stored in the codebook.

In the EFR codec, the codebook for quantizing the spectral coefficients is trained from clean speech. That is, the vectors stored therein represent the spectral envelopes of clean speech frames. When the input vector is speech corrupted by additive noise at a moderate level, it seems likely that the quantizer would have a good chance of choosing a codebook vector that is close to the actual speech, without the noise, effectively enhancing the spectral envelope of the input signal.

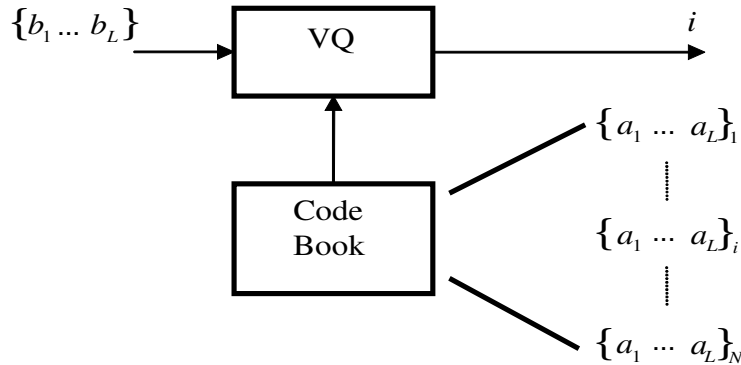


Figure 4.1: The general principle of vector quantization.

4.4 Spectral effects of coding noisy voiced speech

For voiced speech segments, the LP spectrum tends to take the form of fairly defined peaks, centered at the *formant frequencies*, which correspond to resonances in the vocal tract of the speaker. When noise is added to a speech signal, the spectral peaks tend to become flattened and less defined. As formants play a significant role in human speech perception, it is desirable to reproduce the envelope of the LP coefficients as accurately as possible, with emphasis on the peaks around the formant frequencies.

To study the spectral effects of coding a noisy speech signal using the EFR codec, a clean speech signal was corrupted by white noise at 10 dB SNR. The noisy signal was subsequently coded-decoded with EFR. The spectra obtained from 10th order LP analysis of the clean signal, the noisy signal and the noisy signal after coding-decoding with EFR was plotted for three voiced segments. The three voiced segments will be referred to as segments A, B and C, respectively (see figures 4.2, 4.3 and 4.4). Compared to the spectra of the clean signal, the spectra of the noisy signal show flattened peaks. Also, the level is higher because of the added noise. After coding and decoding the noisy signal, however, the signal is spectrally closer to the clean original than the unprocessed noisy signal. The spectra of the processed noisy signal follow the formant peaks of the clean speech much closer, up to about 2.5 kHz. No actual noise suppression was applied, yet the spectra are enhanced. The result is an improvement of the perceived signal quality which is clearly audible. This could be exploited in a decoder based noise suppression system. Less suppression would have to be applied in segments where the coding would already have enhanced the signal. This could lead

to less artifacts being introduced in the processed signal.

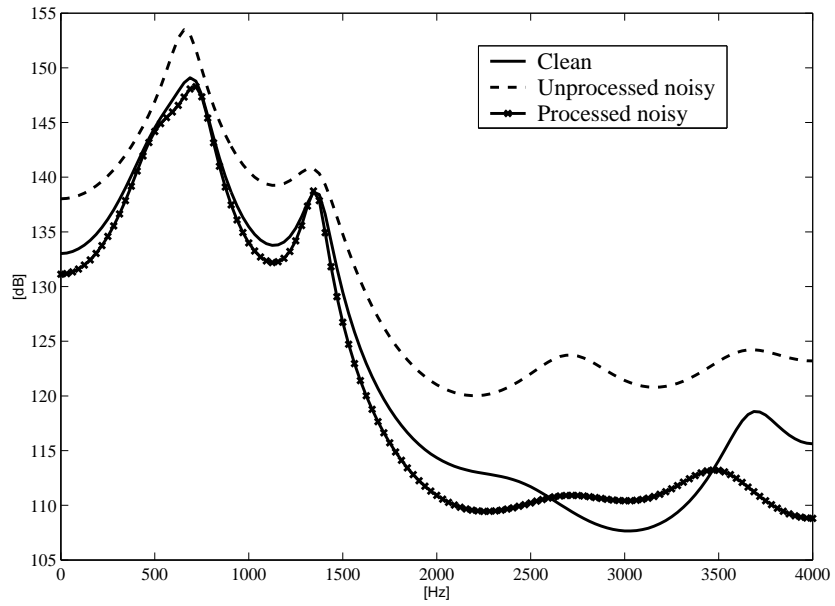


Figure 4.2: LP spectrum of voiced segment A.

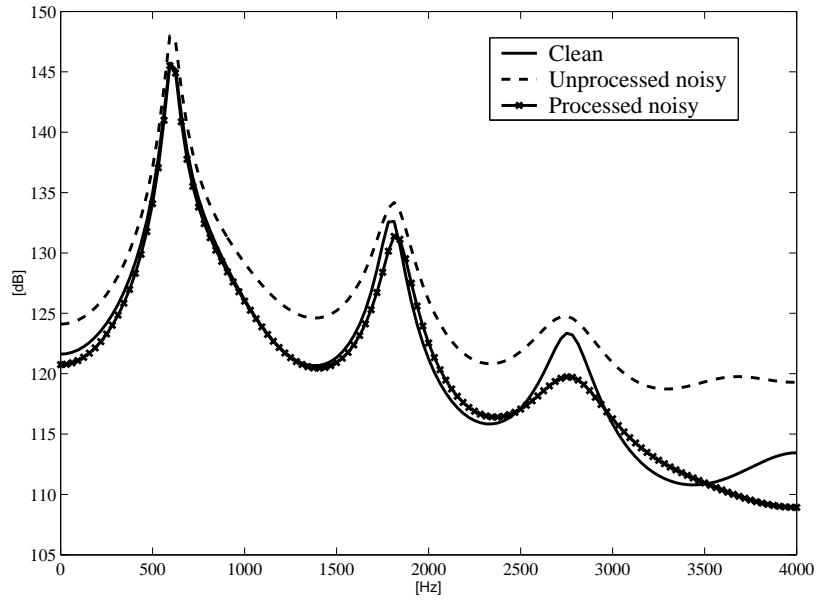


Figure 4.3: LP spectrum of voiced segment B.

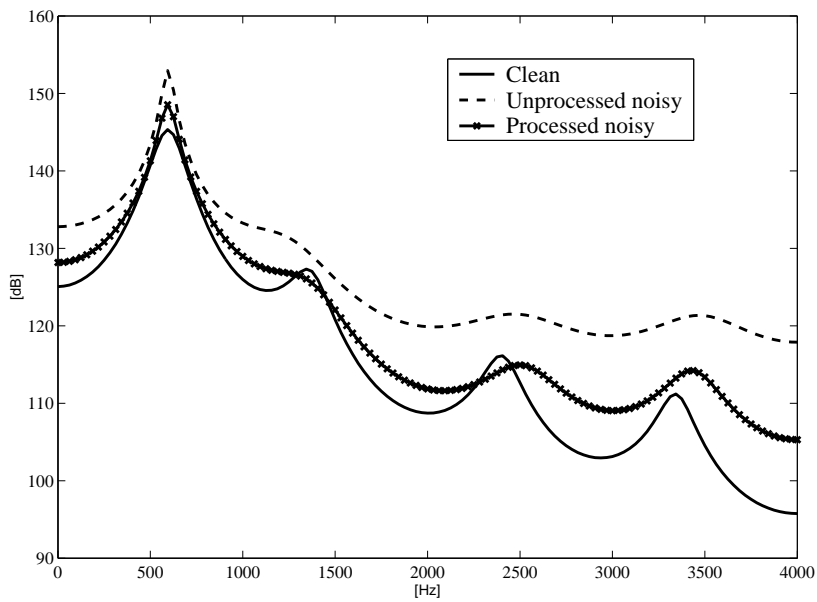


Figure 4.4: LP spectrum of voiced segment C.

Chapter 5

The EVRC noise suppression system

The EVRC NS system uses a frequency domain subtractive scheme to enhance the input speech signal. The diagram below shows the general principle of such a system.

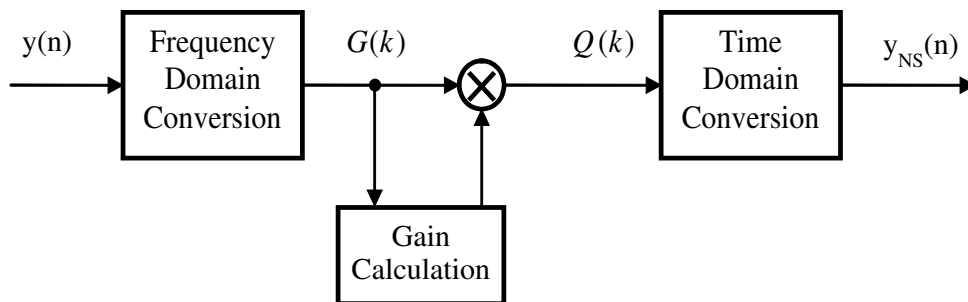


Figure 5.1: General principle of the EVRC NS system.

Firstly, the input signal, $y(n)$, is block wise transformed from the time domain, to the frequency domain. Secondly, a set of gain factors, $q(k)$, are calculated. The actual spectral subtraction takes the form of a multiplication of $G(k)$ with the gain factors from the gain calculation, resulting in the enhanced spectrum $Q(k)$. Lastly, this spectrum is transformed from the frequency domain, to the time domain, and the signal is block wise reassembled to form the enhanced output $y_{NS}(n)$. The three following sections are aimed at more detailed descriptions of how, respectively, the initial time domain processing, frequency domain processing and final time domain processing are implemented in the EVRC NS system.

5.1 Initial time domain processing

The main structure of this part of the NS system is depicted in the diagram below.

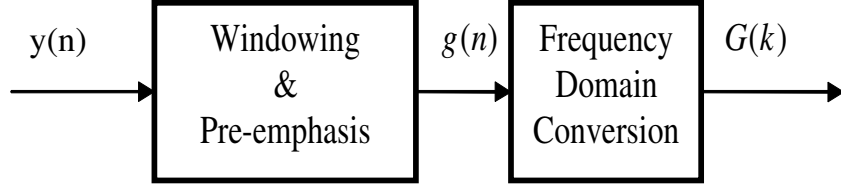


Figure 5.2: EVRC NS initial time domain processing block.

5.1.1 Pre-emphasis

The system operates on blocks of 80 samples (frames), corresponding to 10 ms of speech at 8 kHz sampling rate. The samples are overlapped such that the last $D = 24$ samples of the previous frame make up the D first samples of the input buffer. The next $L = 80$ samples are then pre-emphasized according to

$$d(m, D + n) = y(n) + \varsigma_p y(n - 1); \quad 0 \leq n \leq L \quad (5.1)$$

where ς_p is the pre-emphasis factor, $y(n)$ is the speech input, m is the current frame index and d is the input data buffer (see figure 5.3).

5.1.2 Windowing

To obtain the DFT data buffer g the input data buffer d is windowed with a smoothed trapezoid window according to

$$g(n) = \begin{cases} d(m, n) \sin^2(\pi(n + 0.5)/2D) & 0 \leq n < D \\ d(m, n) & D \leq n < L \\ d(m, n) \sin^2(\pi(n - L + D + 0.5)/2D) & L \leq n < D + L \\ 0 & D + L \leq n < M \end{cases}, \quad (5.2)$$

where M is the DFT sequence length.

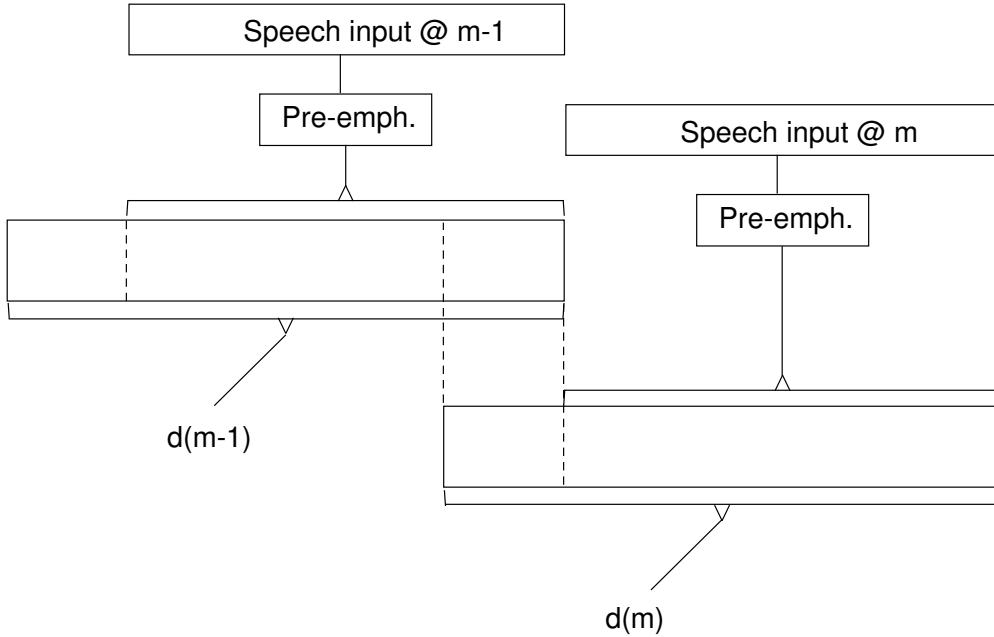


Figure 5.3: EVRC-NS input buffer overlapping and pre-emphasis.

5.1.3 Frequency domain conversion

The buffer $g(n)$ is transformed to the frequency domain,

$$G(k) = R_FFT \{g(n)\}, \quad (5.3)$$

where $R_FFT \{\cdot\}$ denotes a decimation-in-time implementation of the FFT algorithm. This algorithm forms an $M/2$ point complex sequence from an M point real input sequence.

5.2 Frequency domain processing

Now that the input frame has been transformed to the time domain, the stage is set for the actual noise suppression signal processing. The structure of this part is shown in the diagram below. Taking as its input the frequency domain sequence from the previous section, the **channel energy estimator** divides this spectrum into N_c channels and calculates an estimate of the signal energy in each one. The **spectral deviation estimator** calculates the difference between the current channel energies and an average long-term estimate. An estimated signal-to-noise ration is calculated by the **SNR**

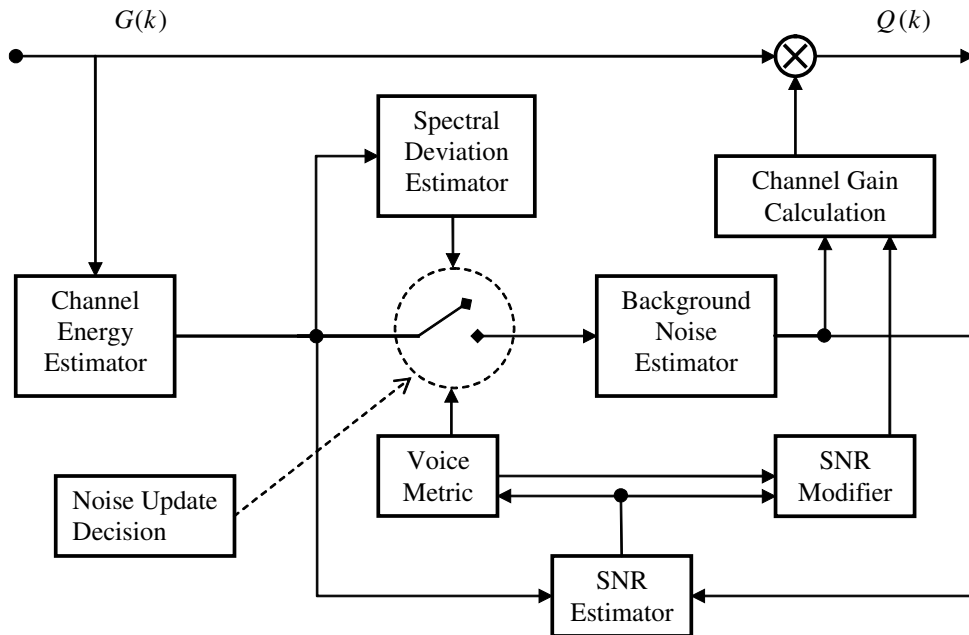


Figure 5.4: EVRC NS frequency domain processing block.

estimator, using the channel energy and background noise estimates. The SNR estimate is used to calculate the **voice metric**, which is a weighted sum which provides an estimate of the signal "quality". It is used mainly as an indication as to whether or not the current frame contains speech. When the input signal is deemed to contain no speech, the **background noise estimator** is updated. Under some conditions the SNR estimates are changed by the **SNR modifier**. Based on the (modified) SNR estimates and the background noise the gains for each channel are calculated by the **channel gain calculator**. These gains are then used to perform the filtering of the input signal. The following sub-sections describe the frequency domain processing blocks in more detail.

5.2.1 Channel energy estimator

From the frequency domain sequence corresponding to the current frame, m , the energy estimates for each of the $N_c = 16$ channels are calculated as,

$$E_{ch}(m, i) = \max \left\{ E_{min}, \alpha_{ch} E_{ch}(m-1, i) + (1 - \alpha_{ch}) \frac{1}{f_H(i) - f_L(i)} \sum_{k=f_L(i)}^{f_H(i)} |G(k)|^2 \right\} \\ 0 \leq i < N_c, \quad (5.4)$$

where i is the channel index, $f_L(i)$ and $f_H(i)$ denote the i^{th} elements from the low and high channel combining tables, respectively (see Appendix). These tables define the division of the $M/2$ point spectrum into N_c channels. This grouping reduces the complexity of the processing. E_{min} is the minimum allowable channel energy and α_{ch} is the channel energy smoothing factor. For the first frame, α_{ch} is set to zero, causing the channel energy estimates to be initialized to the unfiltered channel energies of the first frame, limited downwards by E_{min} . The total channel energy estimate for the current frame is the logarithm of the sum of the channel energy estimates, i.e.

$$E_{tot}(m) = 10 \log_{10} \left(\sum_{i=0}^{N_c-1} E_{ch}(m, i) \right). \quad (5.5)$$

5.2.2 Channel SNR estimator

From the channel energy estimates, and the current channel noise energy estimate, $E_n(m, i)$ (see section 5.2.8), the quantized channel SNR indices $\sigma(i)$ are estimated according to

$$\sigma(i) = \max \left\{ 0, \min \left\{ 89, \text{round} \left\{ 10 \log_{10} \left(\frac{E_{ch}(m, i)}{K_\sigma E_n(m, i)} \right) \right\} \right\} \right\}, \quad (5.6) \\ 0 \leq i < N_c$$

where $K_\sigma = 0.375$ is a scaling constant and $\text{round}\{\cdot\}$ denotes rounding to nearest integer.

5.2.3 Voice metric calculation

The quantized channel SNR indices are used to form an estimate of the signal "quality" called the sum of voice metrics.

$$v(m) = \sum_{i=0}^{N_c-1} V(\sigma(i)), \quad (5.7)$$

where $V(k)$ is the k^{th} entry of the voice metric table \mathbf{V} , as defined in the Appendix.

5.2.4 Spectral deviation and power spectral estimator

The estimated spectral deviation between the current power spectrum $E_{dB}(m, i)$ and the average long-term power spectral estimate $\overline{E_{dB}}(m, i)$ is calculated as

$$\Delta_E(m) = \sum_{i=0}^{N_c-1} |E_{dB}(m, i) - \overline{E_{dB}}(m, i)|, \quad (5.8)$$

where

$$E_{dB}(m, i) = 10 \log_{10}(E_{ch}(m, i)); \quad 0 \leq i < N_c, \quad (5.9)$$

and $\overline{E_{dB}}(m, i)$ is the long-term power spectral estimate calculated at the time of the previous frame, except for the first frame ($m = 1$), for which

$$\overline{E_{dB}}(m, i) = E_{dB}(m, i); \quad 0 \leq i < N_c. \quad (5.10)$$

Next, the long term power spectral estimate is updated for the next frame.

$$\begin{aligned} \overline{E_{dB}}(m+1, i) &= \alpha(m)\overline{E_{dB}}(m, i) + (1 - \alpha(m))E_{dB}(m, i) \\ &0 \leq i < N_c \end{aligned} \quad (5.11)$$

where the exponential windowing factor $0.50 \leq \alpha(m) \leq 0.99$ is a function of $E_{tot}(m)$, such that $\alpha(m)$ grows linearly from 0.5 to 0.99 as $E_{tot}(m)$ goes from 30 to 50 dB (See Appendix for exact equation).

5.2.5 SNR estimate modifier

Before calculating the channel gains the channel SNR estimates are processed further and for each channel a decision is made whether or not to modify the SNR estimate. The result is limited downwards to the SNR threshold σ_{th} , and is denoted $\sigma''(i)$. See Appendix for a detailed pseudo-code description of this process.

5.2.6 Channel gain calculation

The overall gain factor for the current frame, γ_n , is calculated according to

$$\gamma_n = \max \left\{ \gamma_{min}, -10 \log_{10} \left(\frac{1}{E_{floor}} \sum_{i=0}^{N_c-1} E_n(m, i) \right) \right\}, \quad (5.12)$$

where γ_{min} is the minimum overall gain, $E_{floor} = 1$ is the noise floor energy and $E_n(m, i)$ is the estimated noise spectrum calculated during the previous frame (see section 5.2.8). The dB-scale channel gains are calculated as

$$\gamma_{dB}(i) = \mu_g(\sigma''(i) - \sigma_{th}) + \gamma_n; \quad 0 \leq i < N_c \quad (5.13)$$

where μ_g is the gain slope and σ_{th} the SNR threshold, both constants. In figure 5.5 the gain curve for a single channel resulting from equation 5.13 is plotted in comparison with the gain curve resulting from the spectral subtraction rule (see equation 2.4). To simulate the single channel behavior of EVRC-NS, $\sigma'' = \max\{\sigma_{th}, \sigma\}$ and $\sum_{i=0}^{N_c-1} E_n(m, i) = N_c \cdot E_n$ was used. As is evident, the gain curve for EVRC-NS is quite different from that of spectral subtraction.

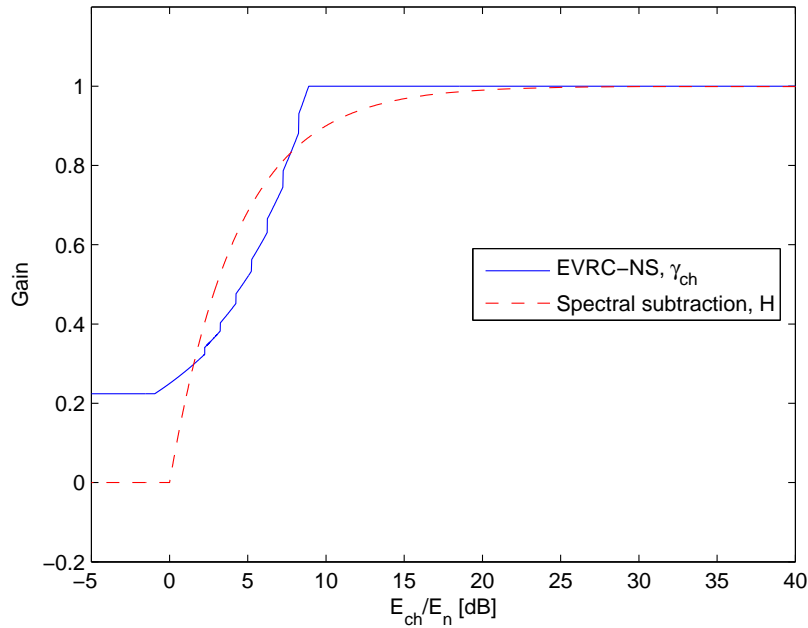


Figure 5.5: Single channel gains for EVRC-NS and Spectral subtraction.

The channel gains are converted to linear scale according to

$$\gamma_{ch}(i) = \min \left\{ 1, 10^{\frac{\gamma_{db}(i)}{20}} \right\} \quad 0 \leq i < N_c \quad (5.14)$$

5.2.7 Frequency domain filtering

The linear channel gains are applied to the sequence $G(k)$ to perform the frequency domain filtering, resulting in the enhanced sequence $Q(k)$:

$$Q(k) = \begin{cases} \gamma_{ch}(i)G(k); & f_L(i) \leq k \leq f_H(i); & 0 \leq i < N_c \\ G(k); & 0 \leq k < f_L(0); & f_H(N_c - 1) < k \leq M/2 \end{cases} \quad (5.15)$$

Note that the the lowest three frequency bins are not modified. While the reason is not explicitly given in the documentation to this system, it is likely due to the fact that telephone networks limit the bandwidth of speech signals to between 300 and 3400 Hz.

5.2.8 Background noise estimate update

The last step in the frequency domain processing is to make a decision whether to update the noise estimate and, if so, calculate new channel noise estimates. Generally, if the voice metric (see section 5.2.3) is below a certain threshold, or if the total channel energy is larger than 0 dB and the spectral deviation is less than a certain threshold, the noise estimate is updated. For a detailed description of the update decision, in pseudo code, see Appendix. If (and only if) the decision is to update the noise estimate

$$E_n(m+1, i) = \max \{E_{min}, \alpha_n E_n(m, i) + (1 - \alpha_n) E_{ch}(m, i)\}; \quad 0 \leq i < N_c \quad (5.16)$$

where E_{min} is the minimum allowable channel energy, and α_n is the channel noise smoothing factor. For the first four frames, the channel noise estimates are initialized to the corresponding estimated channel energies:

$$E_n(m, i) = \max \{E_{init}, E_{ch}(m, i)\}; \quad 1 \leq m \leq 4; \quad 0 \leq i < N_c, \quad (5.17)$$

where E_{init} is the minimum allowable channel noise initialization energy.

5.3 Final time domain processing

This part of the system, which is the inverse of the initial time domain processing, is depicted in figure 5.3. Here the enhanced frequency domain sequence is transformed back to the time domain, and the signal frames are reassembled to form the output signal.

5.3.1 Time domain conversion

The enhanced sequence is converted to the time domain:

$$q(m, n) = R_FFT^{-1} \{Q(k)\}, \quad (5.18)$$

where $R_FFT^{-1} \{\cdot\}$ denotes the inverse of the FFT described in section 5.1.3.

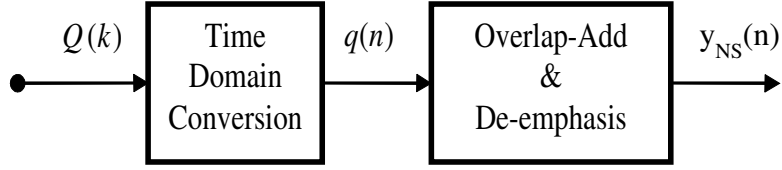


Figure 5.6: EVRC NS final time domain processing block.

5.3.2 Overlap-add and de-emphasis

The output buffer, $y_{NS}(n)$, is formed by overlap-adding the current frame with the previous according to

$$q'(n) = \begin{cases} q(m, n) + q(m-1, n+L); & 0 \leq n < M-L \\ q(m, n); & M-L \leq n < L \end{cases}, \quad (5.19)$$

and applying de-emphasis to $q'(n)$ as

$$y_{NS}(n) = q'(n) + \varsigma_d y_{NS}(n-1); \quad 0 \leq n < L, \quad (5.20)$$

where ς_d is the de-emphasis factor.

Chapter 6

User controlled noise suppression

As the focus of this thesis was on exploring decoder based noise suppression in the context of real-world systems, the noise suppression system was chosen accordingly. The EVRC NS provides a tried and tested commercial system with parameters that control several aspects of its function.

6.1 Controlling the EVRC noise suppression system

After studying the EVRC NS system, three parameters were chosen as potential candidates for user control. These were

- The channel energy smoothing factor, α_{ch} (see equation 5.4).
- The minimum overall gain, γ_{min} (see equation 5.12).
- The channel noise smoothing factor, α_n (see equation 5.2.8).

The two smoothing factors, α_{ch} and α_n , affect how closely the system follows changes in the input signal. When setting smoothing parameters such as these there is a trade-off between optimum performance for slowly (nearly stationary) and quickly changing (non-stationary) signals. Giving the user control allows him/her to adjust the system to the conditions at hand.

The minimum overall gain, γ_{min} , directly affects the channel gains in such a way that it determines how much of the estimated noise energy that the system subtracts from the input signal. In spectral subtraction noise suppression systems such as this, the amount of artifacts introduced in the processed

signal tends to increase with the magnitude of the subtraction. User control enables the listener to adjust the trade-off between artifacts/distortion and background noise level to his/her preference. User control over these three parameters was implemented and evaluated in listening tests, the results and setup of which are presented in chapter 7.

Chapter 7

Evaluation

7.1 The effect of coding a noisy speech signal

To objectively evaluate the effect of coding a noisy speech signal, the method Perceptual Evaluation of Speech Quality (PESQ) [19], was used. This method is a recommendation from ITU-T for predicting the subjective perceived quality of telephony and narrow-band speech signals. It includes an ANSI-C reference implementation. The output quality measure is mapped to the Mean Opinion Score (MOS) scale (see table 7.1), which is the opinion scale recommended by ITU-T for listening quality tests [20]. While PESQ cannot be used to replace subjective listening tests, the correlations between PESQ MOS scores and those from listening tests have been found to be better than 90%.

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

Table 7.1: The opinion scale recommended by ITU-T for listening quality tests.

To evaluate a signal using PESQ, two inputs are needed: the signal to be evaluated, here referred to as the test signal, and a reference signal. The method is best illustrated with an example. Say that the performance of a transmission system is to be evaluated. The test signal would then be produced by passing a clean speech signal through the system, while the clean,

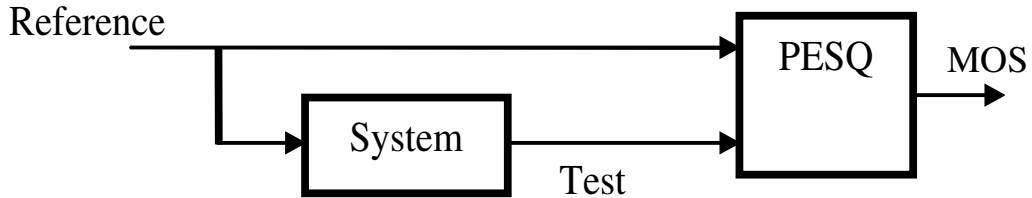


Figure 7.1: Evaluating a system with PESQ using clean speech input.

unprocessed signal would be the reference, as in figure 7.1. When testing with noisy input, the recommended setup is as in figure 7.2. Three series

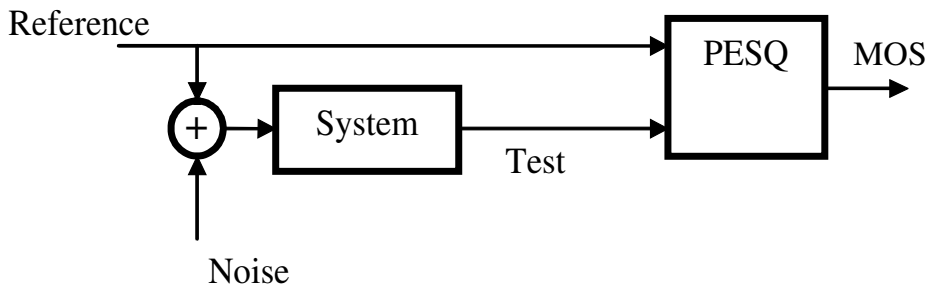


Figure 7.2: Evaluating a system with PESQ using noisy speech input.

of PESQ evaluations were performed, one for each of the three noises used. These noises were taken from Noisex-92 [21] and included white noise, machine noise recorded in a factory and a babble noise consisting of several simultaneous speakers. The three noises will be referred to as white, factory, and babble noise, respectively. The reference signal consisted of four concatenated clean sentences from the TIMIT speech corpus [22], two male and two female speakers. For these tests all signals were downsampled to 8 kHz sampling rate. To the sentences was added noise at the following signal to noise ratios: 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. The quality of these noisy speech signals was evaluated with PESQ using a setup as in figure 7.3. The noisy speech signals were then coded and decoded using the AMR codec at the following bit rates: 12.2 kbits/s, 10.2 kbits/s, 7.95 kbits/s and 6.7 kbits/s. The quality of these processed noisy signals was evaluated as in figure 7.4

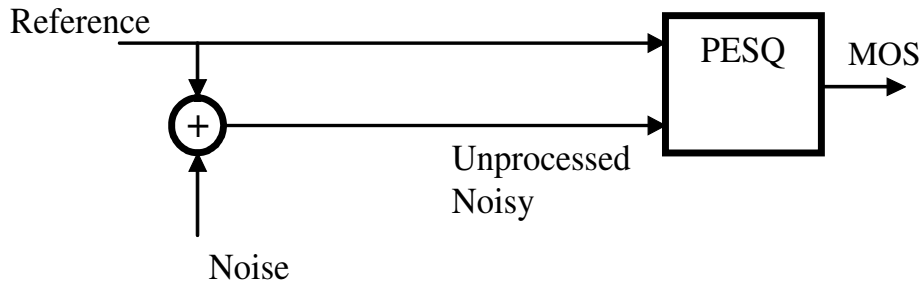


Figure 7.3: The setup used for evaluating the signal quality of the noisy speech signals.

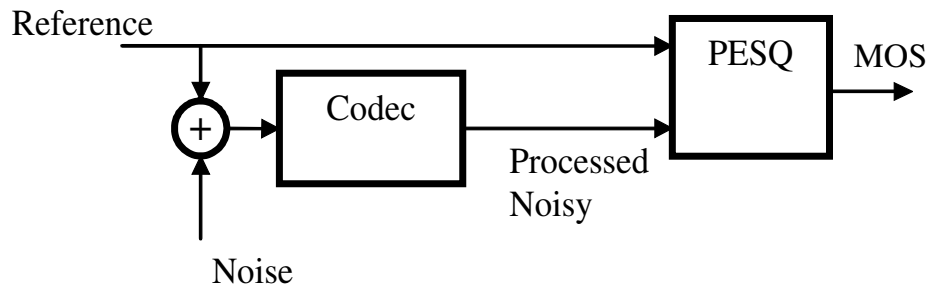


Figure 7.4: The setup used for evaluating the signal quality of the processed (coded-decoded) noisy speech signals.

7.1.1 Post coder noise suppression

For decoder based noise suppression to be feasible, it is of course necessary that the noise suppression system can function in a post processor configuration. In the preliminary studies for this thesis, it was clear that the EVRC noise suppression system performed well as a post processor to the EFR codec. To quantify this, a PESQ evaluation using the setup in figure 7.5 was performed. White noise was added at 0 dB, 5 db, 10 dB, 15 dB and 20 dB SNR. The reference signals were the same as those described in section 7.1.

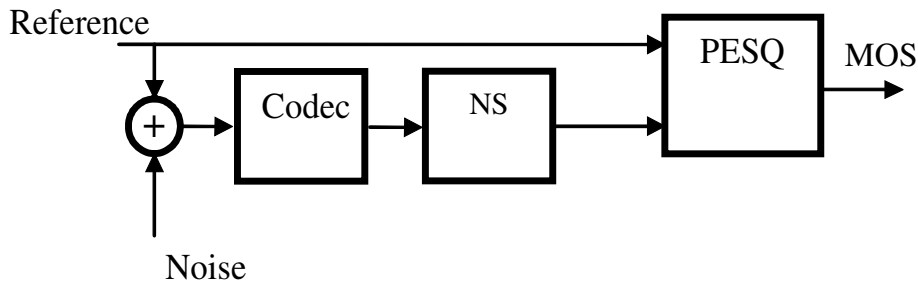


Figure 7.5: The setup used for evaluating post coder noise suppression.

7.2 User controlled noise suppression

A listening test to evaluate user controlled noise suppression was also performed. The setup of the test is depicted in figure 7.6. The test signals used consisted of short sentences from two speakers (one male, one female) from the TIMIT speech corpus contaminated with noises that included white noise and babble noise, both taken from Noisex-92, as well as a distinctly non-stationary rain noise. The signals were processed with the EFR codec prior to NS processing, thereby simulating a decoder based NS system.

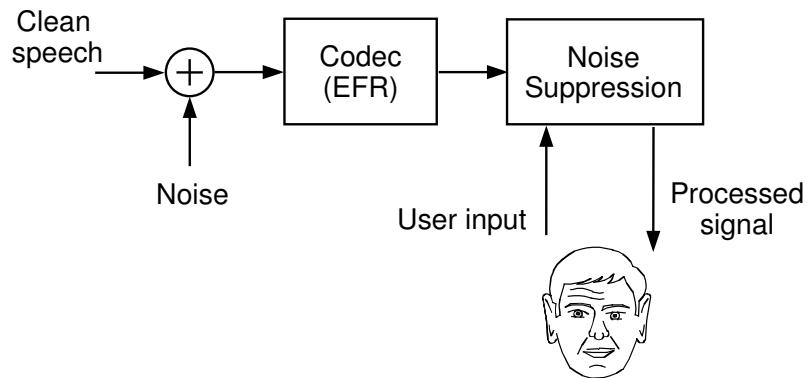


Figure 7.6: The setup of the user control listening test.

An application was simulated with a graphical user interface (GUI). The subjects were presented with an interface with a slider control as in figure 7.7. The slider was set to control one parameter of the EVRC noise suppression system at a time. There were three configurations of the test GUI. The slider

either controlled the settings for the channel energy smoothing factor, α_{ch} [0..1], the channel noise smoothing factor, α_n [0..1] or the minimum overall gain, γ_{min} [-21.. - 5].

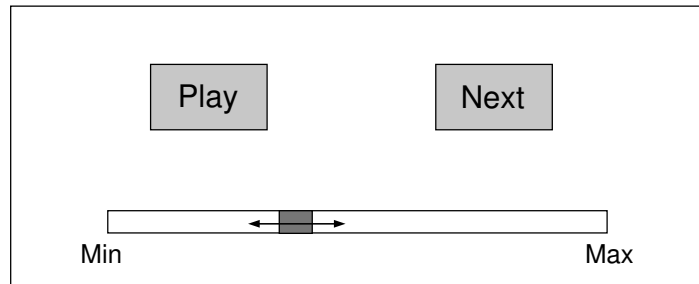


Figure 7.7: The graphical user interface of the test software used to evaluate user control over EVRC noise suppression.

There were two buttons in the interface. When the subject clicked on the button labelled "Play", EVRC noise suppression was applied to the noisy sentence, using the setting from the slider control, and the result was played to the listener via headphones. Ten experienced listeners participated in the test. The subjects were asked to listen and adjust the setting to their preference for each presented sentence. When the subjects clicked the button labelled "Next", the setting of the slider was recorded, the next signal was loaded and the subjects were again asked to listen and adjust the setting of the slider.

This method allows the subject to test any number of settings of the given parameter of the noise suppression for each test sentence. The subject can make comparisons between settings and make fine adjustments to arrive at his/her preference. This differs in many ways from a more conventional blind test with a set of pre-defined test signals. The subject can learn what to expect from listening and adjusting the slider and, as the test progresses, the subject will likely be able to find his/her preferred setting more quickly. In the case of pre-defined signals, one has to either attempt to iteratively bracket the subject's preferred setting, adjusting the given parameter up/down and having the subject rate if the result is better or worse, or test the whole range of settings for every sentence. With these methods one would likely need to use a relatively coarse division of the parameter range to keep the test signals to a reasonable number and avoid fatigue in the subjects. The method used in this thesis places higher demands on the subjects, though.

The results of the test are only relevant if they can be trusted to explore the effect of adjusting the system to the extent and that they arrive at their personal preference accurately for each test sentence. From this follows that the subjects should have some experience with evaluating speech signals and the number of sentences should be limited. Also, there is a possibility that psychological factors inherent in the procedure might bias the result. It has been found [23] that people tend to act in a way that is (or that they believe will be perceived as) consistent with previous actions. Since the subjects are aware of what parameter slider settings they have registered as their personal preference, they might be reluctant to register a subsequent setting that differs much from these. This should be weighed against the advantages of having a procedure that incorporates learning.

7.3 Results

The results of the PESQ evaluations are presented in tables 7.2, 7.3, 7.4, and in figures 7.8, 7.9, 7.10, 7.11. The processed noisy signals consistently scored higher than the original noisy signal. The difference was less for babble noise than for white- and factory noise. There was relatively little difference in scores between the different coded-decoded signals, for any given noise type. This suggests that the coding of these noisy signals has had an enhancing effect, the magnitude of which is more dependent on the noise type than on the bit rate of the codec. Applying EVRC noise suppression after the signal has been processed with the EFR codec further improves the score, which suggests that the EVRC NS system does function in this post processor configuration. These results are congruent with what one would expect after having listened to the signals (see section 4.4).

The results of the user control listening tests are presented in figures 7.12, 7.13, 7.16, 7.17, 7.14, 7.15 and in tables 7.5, 7.6 and 7.3.

In figures 7.18, 7.19 and 7.20, the registered preferred setting is presented separately for each subject.

For all three parameters there is a large difference between individuals for each combination of noise, SNR and speaker. Of the ten test participants, subject number three stands out as one whose preferred setting tended to differ significantly from that of the other subjects. Typically, each individual also has different preferences for different conditions. In many cases, listener preferences differ significantly from the original values for the parameters. This seems to suggest that user control over noise suppression is desirable.

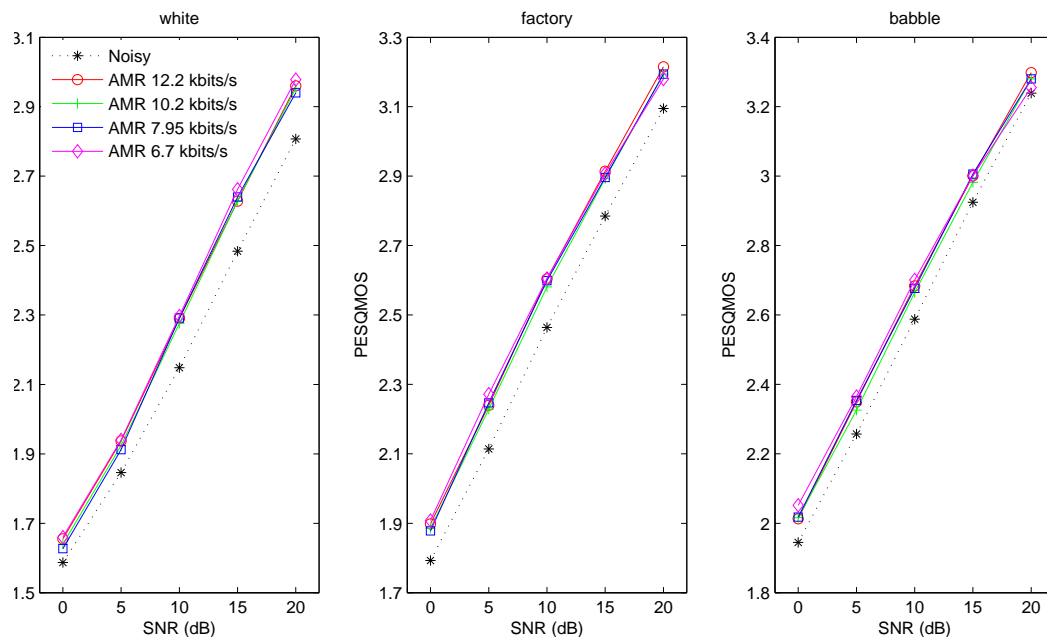


Figure 7.8: PESQ results for white, factory and babble noise.

	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.587	1.846	2.148	2.484	2.807
AMR 12.2	1.655	1.937	2.290	2.628	2.960
AMR 10.2	1.640	1.924	2.276	2.629	2.953
AMR 7.95	1.627	1.912	2.291	2.640	2.940
AMR 6.7	1.660	1.940	2.297	2.662	2.978

Table 7.2: PESQMOS scores for noisy speech and coded-decoded noisy speech. White noise added.

	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.793	2.114	2.464	2.785	3.095
AMR 12.2	1.898	2.240	2.604	2.914	3.215
AMR 10.2	1.886	2.229	2.581	2.893	3.194
AMR 7.95	1.878	2.247	2.599	2.896	3.194
AMR 6.7	1.908	2.272	2.605	2.907	3.180

Table 7.3: PESQMOS scores for noisy speech and coded-decoded noisy speech. Factory noise added.

	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.945	2.257	2.588	2.925	3.239
AMR 12.2	2.013	2.350	2.684	3.001	3.298
AMR 10.2	2.016	2.326	2.665	2.982	3.280
AMR 7.95	2.018	2.354	2.677	3.006	3.280
AMR 6.7	2.052	2.365	2.701	3.001	3.255

Table 7.4: PESQMOS scores for noisy speech and coded-decoded noisy speech. Babble noise added.

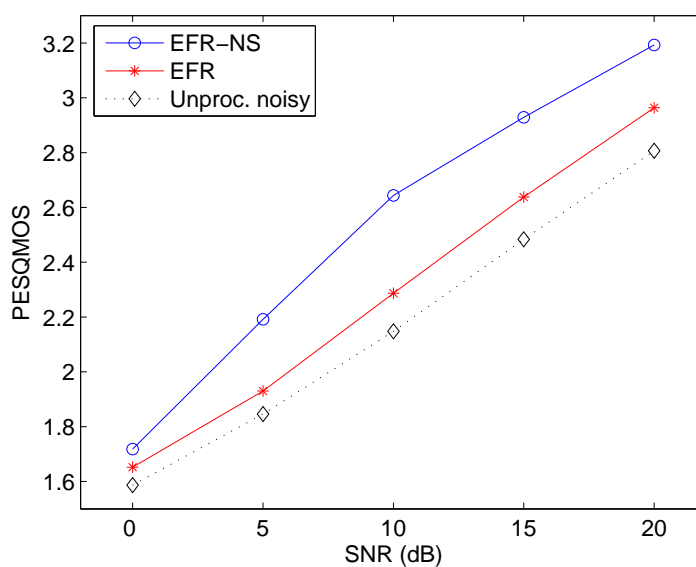


Figure 7.9: PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. White noise added.

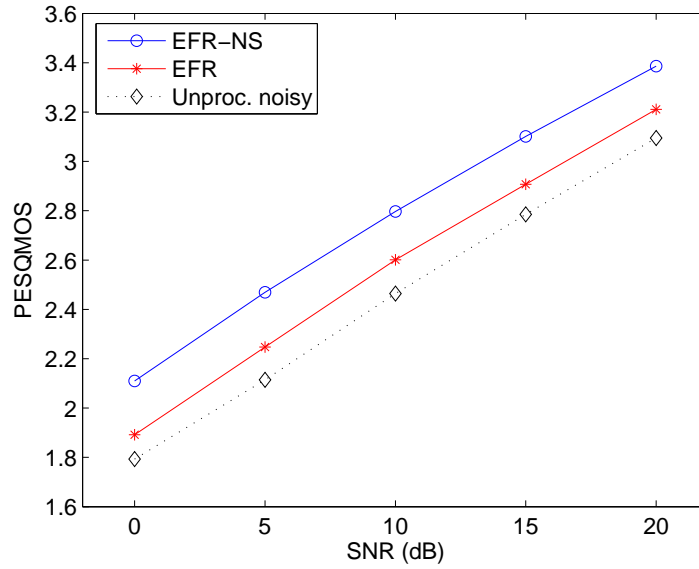


Figure 7.10: PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. Factory noise added.

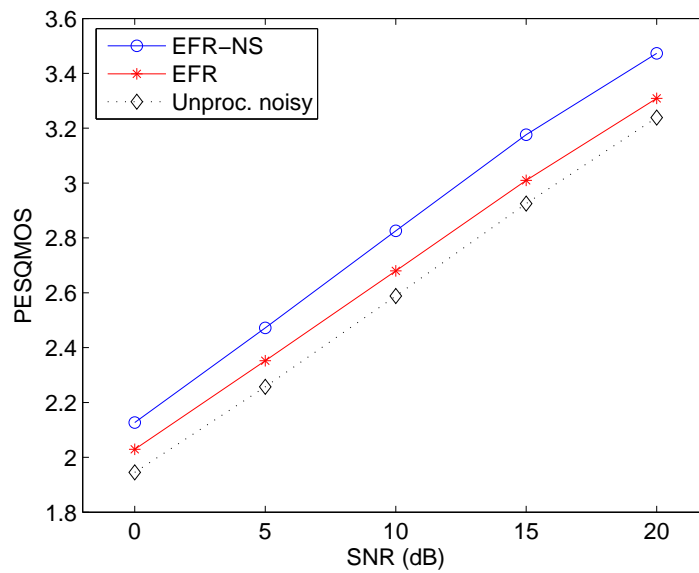


Figure 7.11: PESQMOS scores for noisy speech, coded noisy speech and coded noisy speech post processed with EVRC-NS. Babble noise added.

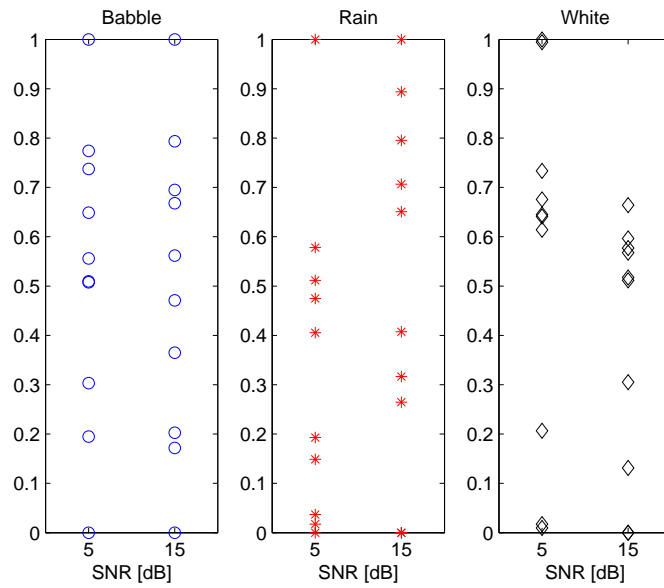


Figure 7.12: Listener preferences for the channel energy smoothing factor, α_{ch} , female speaker.

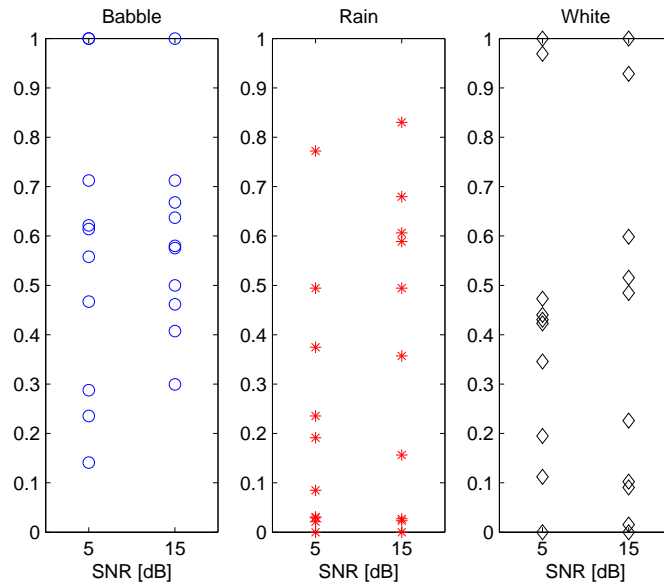


Figure 7.13: Listener preferences for the channel energy smoothing factor, α_{ch} , male speaker.

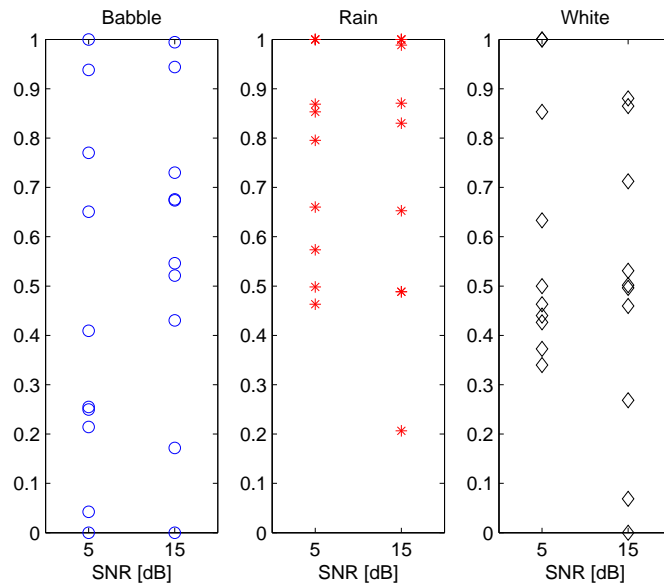


Figure 7.14: Listener preferences for the noise smoothing factor, α_n , female speaker.

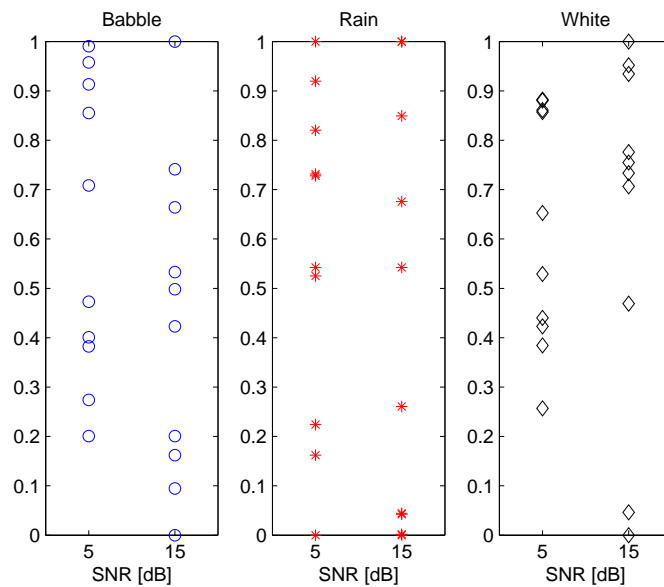


Figure 7.15: Listener preferences for the noise smoothing factor, α_n , male speaker.

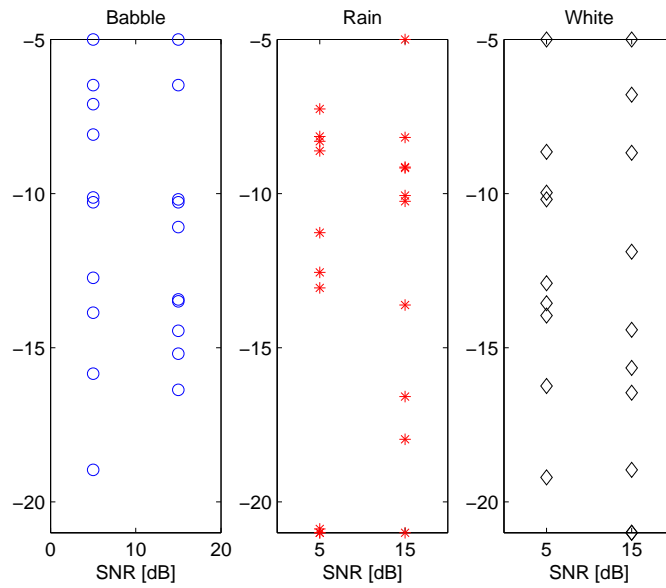


Figure 7.16: Listener preferences for the minimum overall gain, γ_{min} , female speaker.

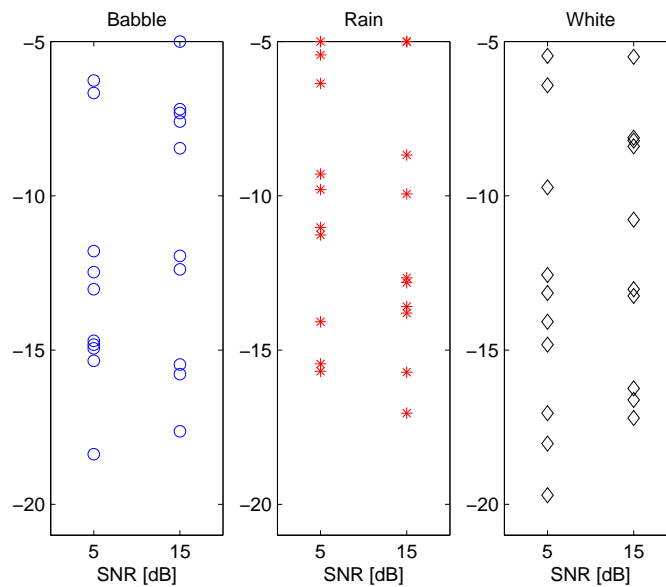


Figure 7.17: Listener preferences for the minimum overall gain, γ_{min} , male speaker.

Speaker	Noise	SNR	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10	Mean	Standard deviation
female	babble	15 dB	0.36486	0	1	0.79344	0.17181	0.69498	0.66795	0.56178	0.2027	0.47104	0.492856	0.311061856
female	babble	5 dB	0.50965	0.19498	1	0	0.50772	0.73745	0.64865	0.55598	0.77413	0.30309	0.523165	0.295024978
female	rain	15 dB	0.40734	0	1	0.70656	0.3166	0.89382	0.65058	0.79537	0.26448	0	0.503475	0.358275088
female	rain	5 dB	0.51158	0.01737	1	0.4749	0.03668	0.40541	0.19305	0.57799	0	0.14865	0.336563	0.318667285
female	white	15 dB	0.30502	0.51158	0	0.56757	0.13127	0.59653	0.57722	0.66409	0	0.51737	0.387065	0.256827631
female	white	5 dB	0.20656	0.01737	1	0.64093	0.67568	0.6139	0.64479	0.73359	0.00965	0.99421	0.553668	0.35965249
male	babble	15 dB	0.5	0.40734	1	0.63707	0.29923	0.71236	0.66795	0.57992	0.57529	0.46139	0.584055	0.192608041
male	babble	5 dB	0.71236	0.14093	1	0.62162	1	0.6139	0.23552	0.46718	0.55792	0.28764	0.563707	0.294521301
male	rain	15 dB	0.60618	0.15637	0.02703	0	0.49421	0.5888	0.67954	0.35714	0.02317	0.83012	0.376256	0.306746228
male	rain	5 dB	0.23552	0.02896	0.08494	0.37452	0.03089	0	0.49421	0.7722	0.02124	0.19112	0.22336	0.254593366
male	white	15 dB	0.59846	0.01544	1	0.48456	0.09073	0.92857	0.51544	0.22587	0	0.10232	0.396139	0.36841542
male	white	5 dB	0.19498	0	1	0.4305	0.96911	0.44015	0.42278	0.47297	0.34556	0.11197	0.438802	0.327477609

Table 7.5: Test subjects' preferences for the channel energy smoothing factor, α_{ch} .

Speaker	Noise	SNR	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10	Mean	Standard deviation
female	babble	15 dB	-10.282	-6.4826	-13.432	-10.189	-16.367	-13.494	-14.452	-11.085	-5	-15.193	-11.59766	3.71860637
female	babble	5 dB	-8.0888	-7.1004	-5	-10.282	-13.865	-10.127	-15.842	-12.734	-6.4826	-18.961	-10.84828	4.461860571
female	rain	15 dB	-10.251	-8.1815	-5	-17.973	-13.618	-16.583	-21	-9.1699	-9.139	-10.066	-12.09814	5.019709085
female	rain	5 dB	-8.305	-8.6139	-21	-20.876	-13.062	-11.27	-21	-12.555	-8.1506	-7.2548	-13.20873	5.682285969
female	white	15 dB	-11.888	-6.7915	-5	-21	-14.421	-16.459	-18.961	-8.6757	-21	-15.656	-13.98522	5.747915383
female	white	5 dB	-8.6448	-5	-5	-16.243	-13.958	-9.973	-19.208	-10.189	-13.556	-12.907	-11.46788	4.604950984
male	babble	15 dB	-8.4595	-7.3166	-5	-12.382	-15.78	-17.633	-11.95	-7.1931	-7.5946	-15.471	-10.87798	4.366657478
male	babble	5 dB	-6.668	-6.2664	-18.375	-12.475	-14.699	-15.347	-14.822	-13.031	-11.795	-14.946	-12.84244	3.819659488
male	rain	15 dB	-12.815	-8.6757	-5	-13.587	-13.803	-5	-12.66	-17.046	-15.718	-9.9421	-11.42468	4.165077495
male	rain	5 dB	-9.8	-5.4324	-5	-14.081	-6.3591	-11.27	-15.687	-15.446	-9.2934	-11.023	-10.33919	3.941556131
male	white	15 dB	-8.3977	-5.4942	-13.031	-17.201	-16.243	-10.776	-16.614	-13.247	-8.1197	-8.2124	-11.7336	4.142644662
male	white	5 dB	-6.4208	-5.4633	-18.035	-17.046	-19.703	-12.568	-14.822	-13.154	-14.081	-9.7259	-13.1019	4.738187056

Table 7.6: Test subjects' preferences for the minimum overall gain, γ_{min} .

Speaker	Noise	SNR	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10	Mean	Standard deviation
female	babble	15 dB	0.52124	0.54633	0.99421	0.72973	0.17181	0.67568	0	0.4305	0.67375	0.94402	0.568727	0.311471099
female	babble	5 dB	0.25483	0.77027	1	0.21429	0.04247	0	0.65058	0.24981	0.93822	0.40927	0.452974	0.363247962
female	rain	15 dB	1	0.98842	1	0.20656	0.48842	1	0.48842	0.65251	0.87066	0.83012	0.752511	0.280265474
female	rain	5 dB	1	0.79537	1	0.86873	0.46332	0.57336	0.49807	0.85328	0.66023	1	0.771236	0.209424296
female	white	15 dB	0.53089	0.71236	0.88031	0	0.45946	0.86486	0.49614	0.06873	0.50193	0.26834	0.478302	0.300078458
female	white	5 dB	0.5	0.85328	1	0.33977	0.37259	0.6332	0.42664	0.44015	0.46332	1	0.602895	0.255797508
male	babble	15 dB	0.74131	0.16216	1	0	0.53282	0.20077	0.49807	0.09459	0.66409	0.42278	0.431659	0.318595294
male	babble	5 dB	0.70849	0.85521	0.99035	0.38224	0.27413	0.20077	0.47297	0.95753	0.91313	0.40077	0.615559	0.302132624
male	rain	15 dB	0	0.26062	1	0.00193	0.84942	0.0444	0.04247	0.67568	0.54247	1	0.441699	0.420720679
male	rain	5 dB	0.5251	0.22394	0.73166	0	0.16216	0.82046	0.54247	0.91969	0.7278	1	0.565328	0.339033713
male	white	15 dB	0	0.04633	0.46911	0.75483	0.95174	0.93436	1	0.70656	0.77606	0.73359	0.637258	0.357380825
male	white	5 dB	0.52896	0.88031	0.85714	0.38417	0.88224	0.25676	0.65251	0.44015	0.42278	0.861	0.616602	0.240110634

Table 7.7: Test subjects' preferences for the noise smoothing factor, α_n .

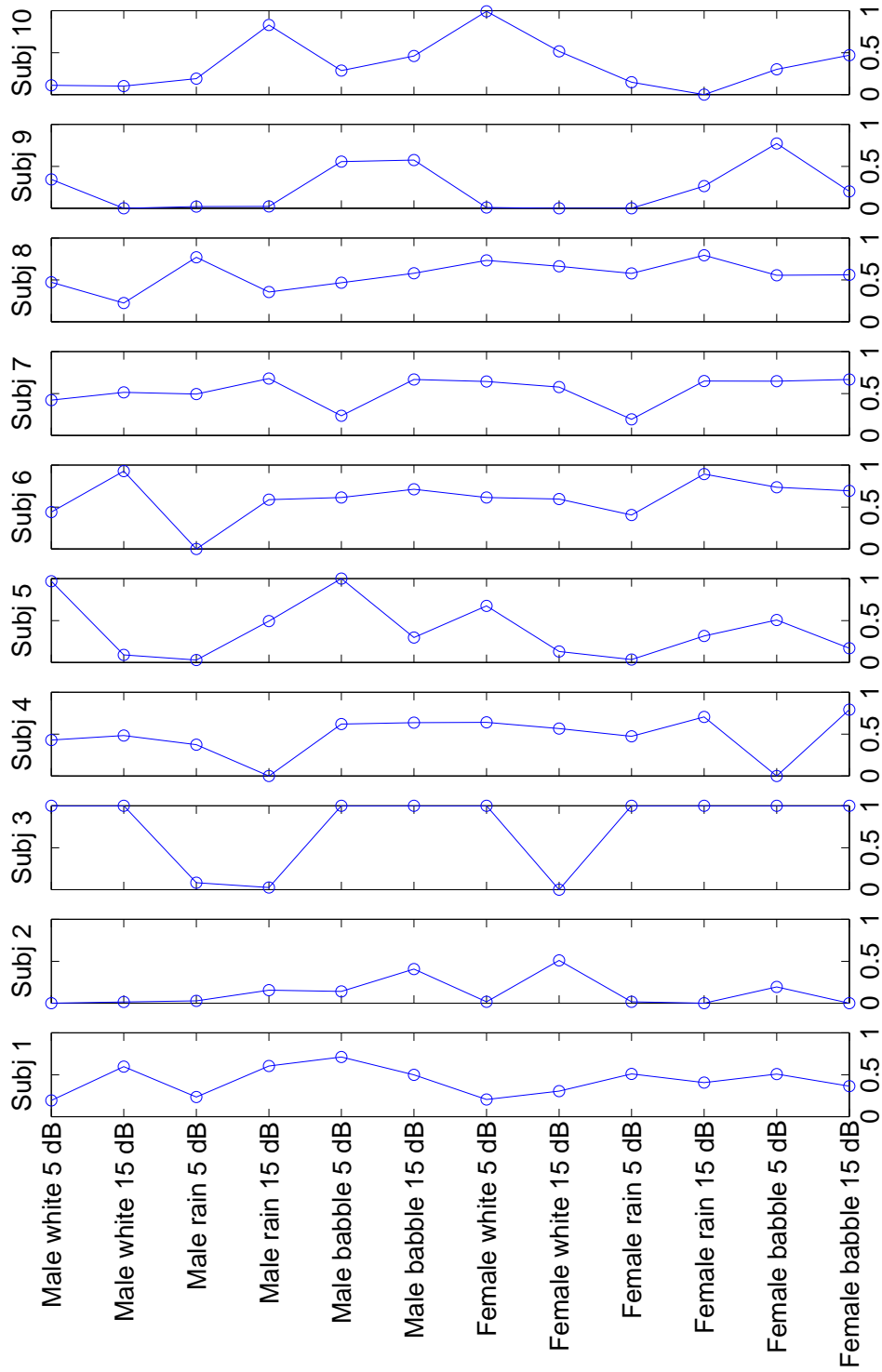


Figure 7.18: Preferences for the channel energy smoothing factor, α_{ch} , for all subjects.

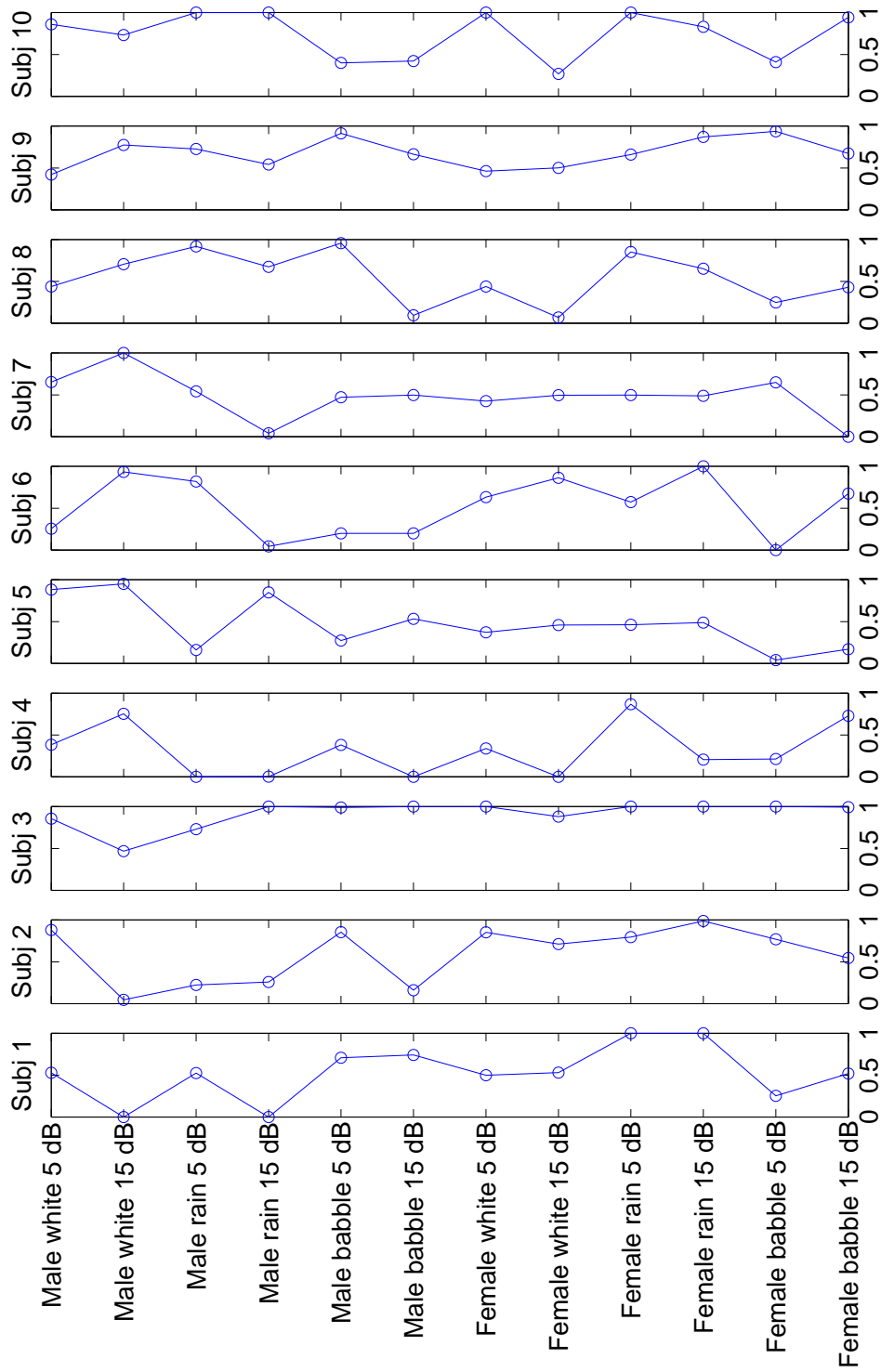


Figure 7.19: Preferences for the noise smoothing factor, α_n , for all subjects.

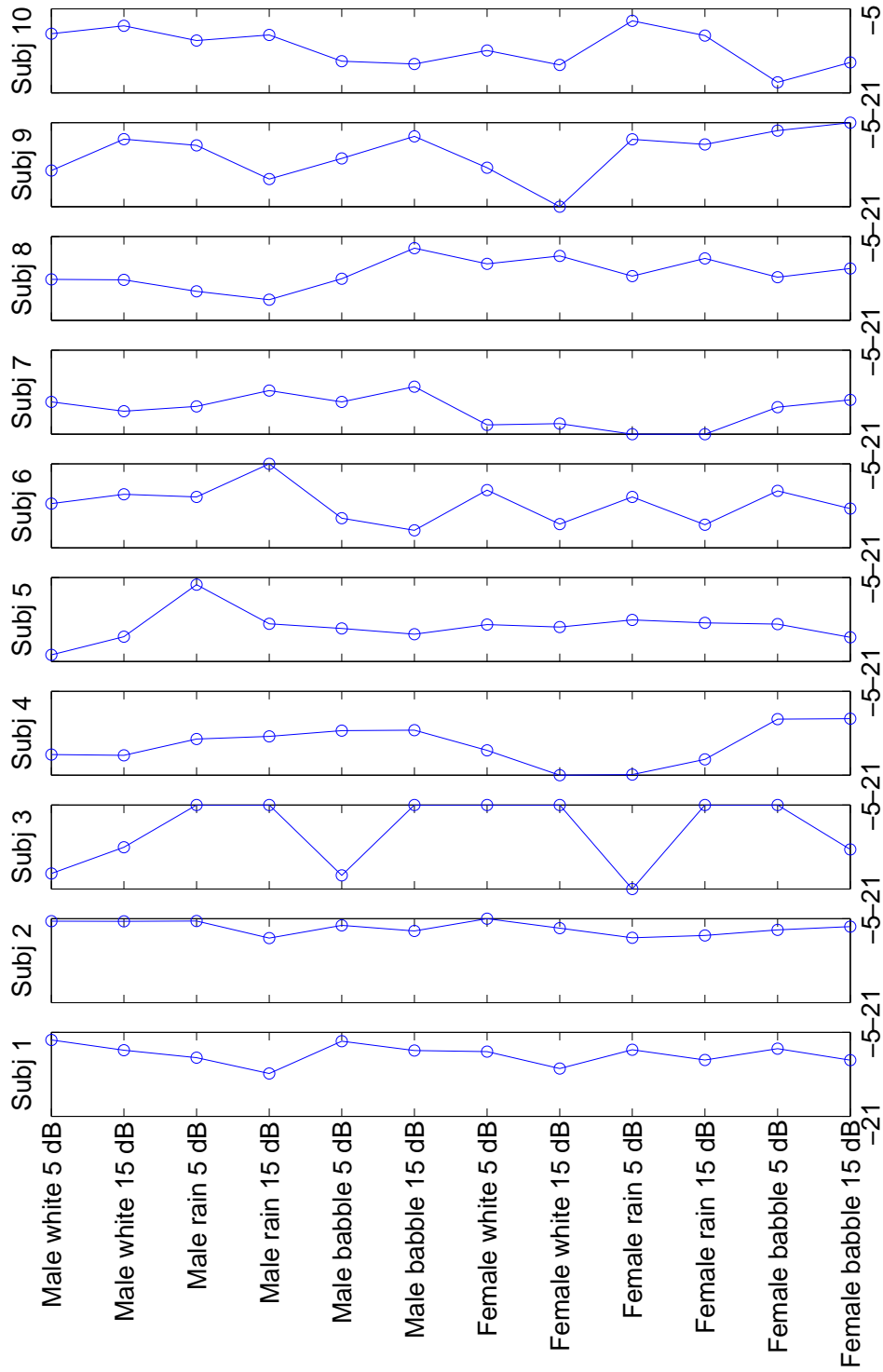


Figure 7.20: Preferences for the minimum overall gain, γ_{min} , for all subjects.

Chapter 8

Conclusions and suggested future work

8.1 Conclusions

Intuitively, a prerequisite for decoder based noise suppression is that the codec used is robust to noisy input. Thus, to explore the feasibility of such a system, the effect of coding a noisy speech signal using CELP-type coders was investigated. The tested codecs (AMR and EFR) proved to be robust to noise and the coding was shown to have a signal enhancing effect. PESQ evaluation confirmed this and indicated that the magnitude of the signal enhancing effect was more dependent on the noise type than on the bit rate of the codec. These results indicate that decoder based noise suppression would be feasible with the tested codecs.

A decoder based noise suppression system would allow user control. After a general study of noise suppression and noise estimation, the subtractive type noise suppression present in the Enhanced Variable Rate Codec (EVRC NS) was studied in detail. This system was used as a post processor to the EFR codec to simulate a decoder based noise suppression system with user control. Three parameters of EVRC NS system were selected for user control and a listening test with ten participants was performed. The test yielded individual preferences for the three parameters given several combinations of speaker, noise type and signal-to-noise ratio. The results show large differences in individual preferences, which would seem to indicate that some sort of user control over noise suppression would be desirable.

8.2 Suggested future work

The signal enhancing effect that coding can have on a noisy speech signal that was indicated in this work would be interesting to explore in more depth. A more thorough study could focus on which parts of the codec contribute and to which extent.

For continued work on user controlled noise suppression, it would be beneficial to design a NS system with specific and clearly defined control parameters with a predictable effect on the system throughout their defined ranges. For a more intuitive testing system, real-time control (during playback) of the NS would be desirable.

Appendix A

EVRC Noise Suppression

A.1 The channel combining tables

$$f_L = \{2, 4, 6, 8, 10, 12, 14, 17, 20, 23, 27, 31, 36, 42, 49, 56\} \quad (\text{A.1})$$

$$f_H = \{3, 5, 7, 9, 11, 13, 16, 19, 22, 26, 30, 35, 41, 48, 55, 63\} \quad (\text{A.2})$$

A.2 The voice metric table

$$\begin{aligned} \mathbf{V} = & \{2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, \\ & 5, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9, 10, 10, 11, 12, 12, 13, 13, \\ & 14, 15, 15, 16, 17, 17, 18, 19, 20, 20, 21, 22, 23, 24, 24, \\ & 25, 26, 27, 28, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 37, 38, \\ & 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 50, 50, 50, 50, \\ & 50, 50, 50, 50, 50\} \end{aligned} \quad (\text{A.3})$$

A.3 The exponential windowing factor, $\alpha(m)$

The exponential windowing factor, $\alpha(m)$, is calculated as a function of total channel energy, $E_{tot}(m)$, according to

$$\alpha(m) = \alpha_H - \frac{\alpha_H - \alpha_L}{E_H - E_L} (E_H - E_{tot}) \quad (\text{A.4})$$

the result is limited to between α_H and α_L by

$$\alpha(m) = \max\{\alpha_L, \min\{\alpha_H, \alpha(m)\}\} \quad (\text{A.5})$$

where E_H and E_L are the energy endpoints (in dB) for the linear interpolation of $E_{tot}(m)$, that is transformed to $\alpha(m)$ which has the limits

$$\alpha_L \leq \alpha(m) \leq \alpha_H. \quad (\text{A.6})$$

The values of these constants are defined as:

$$E_H = 50, E_L = 30, \alpha_H = 0.99, \alpha_L = 0.50. \quad (\text{A.7})$$

A.4 SNR Estimate modification

Determine whether the channel SNR modification should take place, then proceed to modify the appropriate SNR indices:

```

/* Set or reset modify flag */
index_cnt = 0
for ( i = N_m to NUM_CHAN-1 step 1 )
{
    if ( sigma <= INDEX_THLD )
        index_cnt = index_cnt + 1
}
if ( index_cnt < INDEX_CNT_THLD )
    modify_flag = TRUE
else
    modify_flag = FALSE

/* Modify the SNR indices to get sigma_prim */
if ( modify_flag == TRUE )
    for ( i = 0 to NUM_CHAN-1 step 1 )
        if (( v(m) <= METRIC_THLD ) or ( sigma <= SETBACK_THLD ))
            sigma(i) = 1
        else
            sigma_prim(i) = sigma(i)
else
    sigma_prim = sigma

/* Limit sigma_bis to SNR threshold SNR_THLD */
for ( i = 0 to NUM_CHAN-1 step 1 )
    if ( sigma_prim(i) < SNR_THLD )
        sigma_bis(i) = SNR_THLD
    else

```

```
sigma_bis(i) = sigma_prim(i)
```

```
INDEX_THLD=12, INDEX_CNT_THLD=5, METRIC_THLD=45,  
SETBACK_THLD=12, N_m=5, SNR_THLD=6, NUM_CHAN=16.
```

A.5 Noise update decision

The following logic, as shown in pseudo-code, demonstrates how the noise estimate update decision is ultimately made:

```
/* Normal update logic */  
update_flag = FALSE  
if ( v(m) <= UPDATE_THLD )  
{  
    update_flag = TRUE  
    update_cnt = 0  
}  
  
/* Forced update logic */  
else if ( ( E_tot(m) > NOISE_FLOOR_DB ) and ( delta_E < DEV_THLD ) )  
{  
    update_cnt = update_cnt + 1  
    if ( update_cnt <= UPDATE_CNT_THLD )  
        update_flag = TRUE  
}  
  
/* "Hysteresis" logic to prevent long-term creeping of update_cnt */  
if ( update_cnt == last_update_cnt )  
    hyster_cnt = hyster_cnt + 1  
else  
    hyster_cnt = 0  
last_update_cnt = update_cnt  
if ( hyster_cnt > HYSTER_CNT_THLD )  
    update_cnt = 0  
  
UPDATE_THLD=35, NOISE_FLOOR_DB=0, DEV_THLD=28,  
UPDATE_CNT_THLD=50, HYSTER_CNT_THLD=6.
```

Bibliography

- [1] 3rd Generation Partnership Project. "AMR speech CODEC", 3GPP TS 26.071, V6.0.0, December 2004
- [2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April 1979.
- [3] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 2, pp. 137-145, April 1980.
- [4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models", *IEEE Trans. Signal Processing*, vol. 40, pp. 725-735, April 1992.
- [5] J. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition", *IEEE Trans. Signal Processing*, vol. 39, pp. 795-805, April 1991.
- [6] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering", in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, Dallas, USA, April 1987, pp. 177-180.
- [7] J. Gibson, B. Koo and S. Gray, "Filtering of colored noise for speech enhancement and coding", *IEEE Trans. Signal Processing*, vol. 39, pp. 1732-1742, August 1991.
- [8] B. C. J. Moore, "An introduction to the psychology of hearing", *Academic Press, Fifth edition*, 2003, ISBN 0-12-505628-1.
- [9] V. Stahl, A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering", *IEEE*, 2000.

- [10] H. K. Kim, R. V. Cox and R. C. Rose, "Performance improvement of a Bitstream-based front-end for wireless speech recognition systems", *IEEE Transactions on speech and audio processing*, vol. 10, no. 8, pp. 591-604, November 2002.
- [11] R. Chandran and D. J. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement", *Proc. 43rd IEEE Midwest Symposium on Circuits and Systems*, pp. 10-13, August 2000.
- [12] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26, no. 3, pp. 197-210, June 1978.
- [13] X. Huang, A. Acero, H-W. Hon, "Spoken language processing", *Prentice Hall*, 2001, ISBN 0-13-022616-5.
- [14] European Telecommunications Standards Institute (ETSI), "Enhanced Full Rate (EFR) speech transcoding" GSM 06.60 v. 8.0.1, 1999.
- [15] Telecommunications Industry Association (TIA), "Enhanced Variable Rate Codec, Speech service option 3 for wideband spread spectrum digital systems", TIA/EIA/IS-127, September 1996.
- [16] H. Gustafsson, "Speech enhancement for mobile communications", *Blekinge Inst. Of Tech., Dept. of Telecom. and Sig. Proc.*, 2002, ISBN 91-7295-011-0.
- [17] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, March 1999.
- [18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, December 1984.
- [19] Telecommunication standardization sector of ITU, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs", ITU-T P.862 February 2001.
- [20] Telecommunication standardization sector of ITU, "Methods for subjective determination of transmission quality", ITU-T P.800 August 1996.

- [21] Institute for Perception-TNO, The Netherlands. NOISE-ROM-0 signal 003, 011, 021. February 1990.
- [22] DARPA-TIMIT, "Acoustic-phonetic continuous speech corpus", 1990.
- [23] R. B. Cialdini, "Influence: the psychology of persuasion", New York, 1993, ISBN: 0-688-12816-5.