Single-Channel Suppression of Background Noise in Speech Signals

Seminararbeit aus Algorithmen in Akustik und Computermusik 2, SE

Mikko Roininen
Martin Kirchberger

Betreuung: Franz Zotter

Graz, February 28, 2010

UNIVERSITÄT FÜR MUSIK UND DARSTELLENDE KUNST GRAZ - AUSTRIA

institut für elektronische musik und akustik

Abstract

In this seminar paper, two different approaches concerning single channel noise suppression are investigated: spectral subtraction and gammatone filterbank subband processing. Thereby, the theory behind both methods as well as the important aspects concerning the implementation will be explained. Furthermore, a short summary of the acoustically evaluated performance is given to verify the functionality of both algorithms.

Contents

1 Introduction

Speech noise suppression is a traditional and broadly studied research field. This is due to the vast amount of recorded and transferred speech being used in all kinds of environments, the varying quality of the recording and transferring media, and the importance of speech communication in general. Typical applications of speech denoising include:

- Narrow-band voice communications
- Speech recognition
- Speaker authentication
- Voice-controlled systems
- Speech compression

According to Boll denoising can be carried out by using noise-cancelling microphones, internal modification of the voice processor algorithms to explicitly compensate for the noise, or by pre-processor noise reduction. Chen et al. state the following classification for noise suppression techniques:

- The number of channels available for enhancement; i.e., single channel and multi channel techniques.
- How the noise mixes with speech; i.e., additive noise, multiplicative noise, and convolutive noise.
- Statistical relationship between the noise and speech; i.e., uncorrelated or even independent noise, and correlated noise (such as echo and reverberation).
- How the processing is carried out; i.e., in the time domain or in the frequency domain.

This seminar paper concentrates on the single channel pre-processing case with uncorrelated additive noise. Both frequency domain frame-based and time domain sample-by-sample processing have been implemented for comparison. The case considered here is quite typical, for instance, in mobile communications; there's only one microphone available and the main noise sources are the environment and the transmission channel, which are independent from the speech itself.

The denoising methods presented are based on the classical theories of (Power) Spectral Subtraction and Wiener filtering. These basic methods show well the duality of noise reduction and speech distortion. When denoising a noisy signal, more or less distortion is always added to the output signal. With noise suppression being too harsh - especially in low SNR situations - the distortions or residual noise can be perceptually more annoying than the initial noise or, in the case of algorithm pre-processing, deteriorate the algorithm performance.

2 Spectral Subtraction Method

Spectral subtraction is a method to retrieve the pure signal spectrum $X(f)$ by subtracting a noise spectrum estimate $N(f)$ from the noisy signal spectrum $Y(f)$. The following block diagramm allows a quick overview of the processing steps.

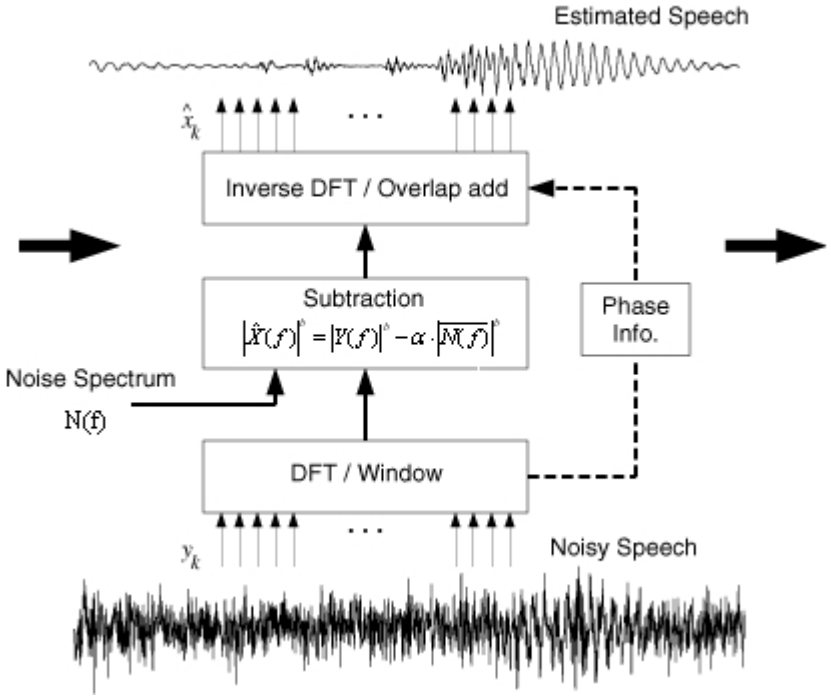

Figure 1: Overview of spectral subtraction method (see *9)

In the following sections, the different stages leading to a suppressed-noise output signal will be explained in detail.

2.1 Estimation of the noise spectrum

For the estimation of the noise spectrum, only the time periods, when the desired signal is absent, are used. It is assumed, that the background noise is a stationary due to a slowly varying process. To differentiate between active and inactive periods, the noisy signal is divided into short consecutive time frames. The calculated energy of the respective time frames is then used as a distinctive feature. With defining a certain threshold, all time frames with lower energy are regarded as pure noise, those with locally increased energy are considered carrying noise and the desired signal. During periods exhibiting desired signal activity, the noise estimation is 'frozen' and the last noise update is kept as an estimation for subsequent signal periods. Before being subtracted from the noisy signal, the noise estimation spectrum is smoothed, i.e. the estimated noise spectrum of preceding time frames are taken into account.

2.2 Comfort noise

Spectral subtraction may not yield negative values for the estimate of the clean, denoised signal. In order to prevent negative values especially at low SNR, a mapping function P has to be employed:

$$P[\hat{X}(f)] = \begin{cases} |\hat{X}(f)| & \text{if } |\hat{X}(f)| > \beta \cdot |Y(f)| \\ \beta \cdot |Y(f)| & \text{if } |\hat{X}(f)| \leqq \beta \cdot |Y(f)| \end{cases} \tag{1}$$

A minimal noise floor $\beta \cdot |Y(f)|$ may be kept as "'comfort"' noise.

2.3 Spectral subtraction

The equation of the generalized spectral subtraction rule is:

$$|\hat{X}(f)|^b = |Y(f)|^b - \alpha \cdot |N(f)|^b \tag{2}$$

Its coefficient α is the (over)subtraction factor, the significance of which will be explained in subsection 2.6.1. The parameter b determines the spectral power of the subtraction. Setting b to the value 1, the equation turns into *magnitude* spectral subtraction. If $b = 2$, the *power* spectrum subtraction rule is applied. Both methods are similar, however, they result in slightly different performances.

2.4 Spectral subtraction as filter function

Instead of subtracting spectral magnitudes, spectral subtraction can be expressed as a filtering of the noisy signal.

$$\begin{gathered} |\hat{X}(f)|^b = |Y(f)|^b - \alpha \cdot |N(f)|^b \\ |\hat{X}(f)|^b = |H(f)| \cdot |Y(f)|^b \\ \text{with } H(f) = 1 - \frac{|N(f)|^b}{|Y(f)|^b} \end{gathered} \tag{3}$$

$H(f)$ is therefore the frequency response of the spectral subtraction filter. It is a zero phase filter, with its magnitude response being within the boundaries of [0,1]. Taking the mapping of subsection 2.2 into account, its range of $H(f)$ diminishes to [β,1]. This property allows another perspective on noise suppression by spectral subtraction: Spectral subtraction is equivalent to a gain function, which attenuates the signal corresponding to its current energy level. To retrieve the time domain signal $\hat{x}(t)$,the magnitude spectrum $|\hat{X}(f)|$ is combined with the phase of the noisy signal before being transformed via the inverse discrete Fourier transform.

$$\hat{x}(t) = \sum_{0 \leq k \leq N-1} |X(k)| \cdot e^{j\Theta(k)} \cdot e^{-\frac{j2\pi}{N} k \cdot m} \tag{4}$$

This restoration is based on the assumption, that the magnitude distortion rather than phase distortion is responsible for the audible noise.

2.5 Artefacts in Spectral Subtraction

The are two main reasons for the annoying distortions:

- inaccuracy of noise estimation assumptions
- nonlinear properties of the mapping

The observed signal distortions can be described as a metallic sound and is often referred to as "'musical noise"'. For a short period of time, narrow bands of frequencies have a small peak in the signal spectrum. If those bands are in the vicinity of the signal, they are masked and therefore inaudible. However, if components around the noise-induced short peak are not present, the outcome is disturbing.

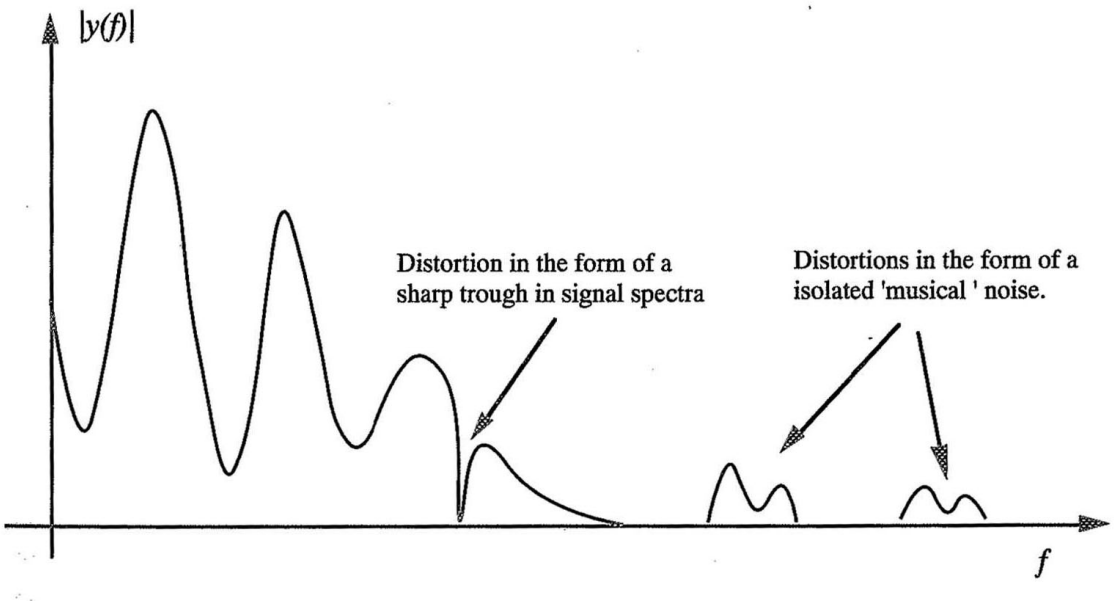

Figure 2: Musical noise (see *1) p.249)

Especially at lower signal-to-noise ratios, simple implementations of the spectral subtraction can lead to even worse signal qualities than the noisy input.

2.6 Counteractive measures

2.6.1 Oversubtraction

One effective counteractive measure against musical noise is oversubtraction. In this case, the (over)subtraction factor α from equation 2 is set to values higher than 1.

Thus, more than the estimated noise is subtracted from the signal. The benefit is to decrease the musical noise as the narrow-band peaks are more likely to fall below the lower limit $\beta \cdot |Y(f)|$. Especially in the case of low SNR, the oversubtraction method is necessary to achieve satisfactory results. Common values for α are between 1 and 2 in the magnitude domain and in the range of 1 and 4 in the power domain, respectively.

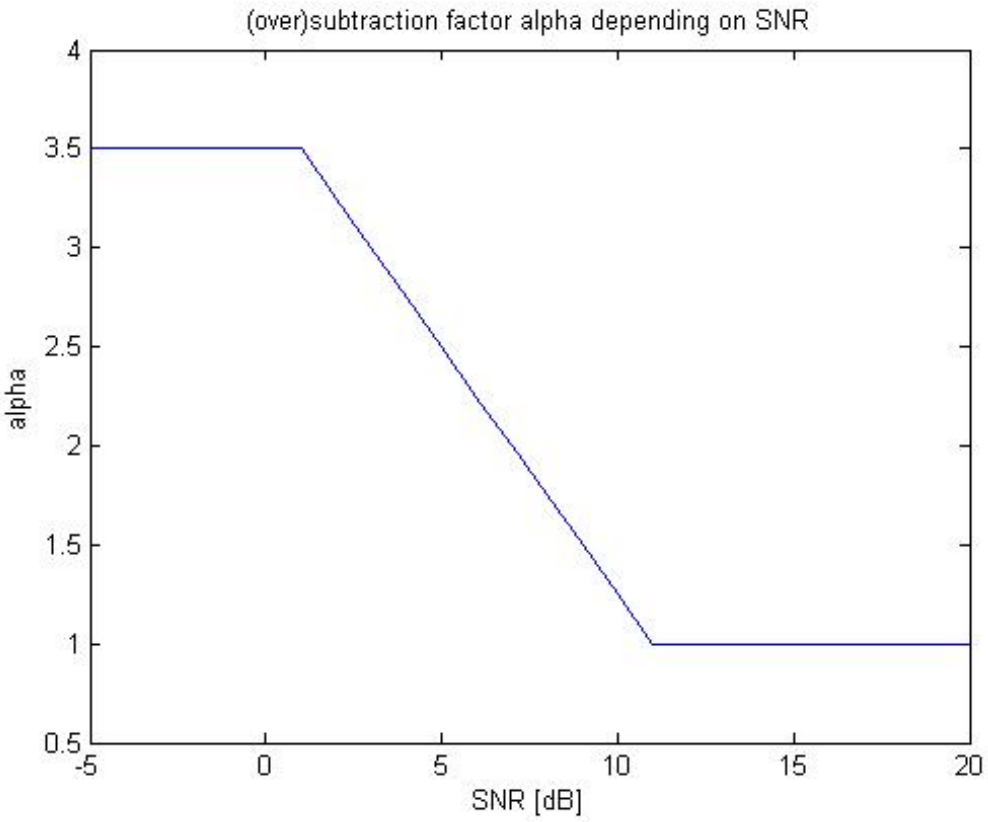

Figure 3: Different values for α depending on the SNR of the power spectra.

2.6.2 Post processing

As stated before, musical noise is a distortion caused by short narrow-band peaks in the spectral domain isolated from the wanted signal. These features can be used to identify and remove a big part of the annoying distortions. Two major properties have to be fulfilled in order to identify musical noise in the frequency domain. First, for musical noise the number of consecutive bins with magnitude above the threshold is usually smaller than a certain maximum, depending on the sampling rate and the block size of the time frames. Second, all those values stay below a certain level. Only if both criteria are met, the content of these bins is regarded as musical noise. To suppress this, an exception is applied to the noise-suppression rule. The respective components are attenuated with β, the lower limit of the gain function. The following figure illustrates

the process.

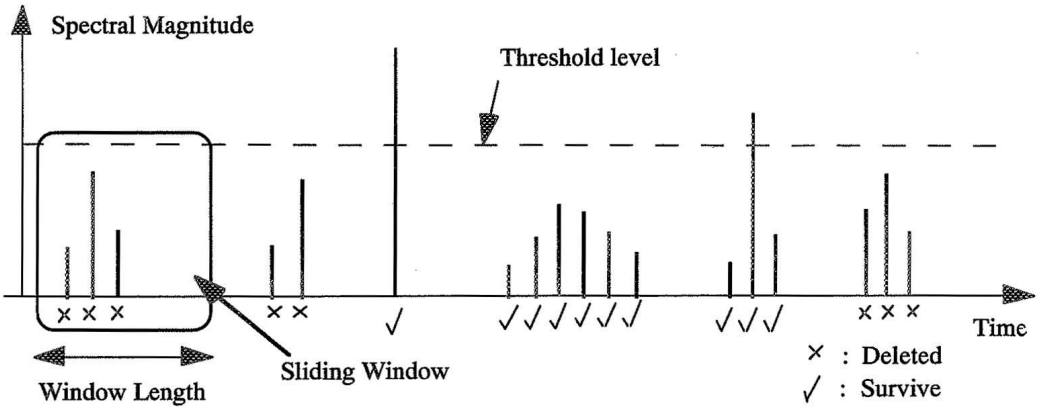

Figure 4: Post processing to identify and delete musical noise (see *1) p.252)

2.7 Summary on Spectral Subtraction

In order to implement a well-working spectral subtraction program, several processing steps have to be considered. As discussed in previous sections, noise estimation is necessary to distinguish between periods of pure noise and periods of noisy signal. A mapping function is introduced to prevent negative values in the signal spectrum estimate. Moreover, counteractive measures have to be attached in order to decrease the disturbance caused by musical noise. Finally, the spectral estimate is combined with the original phase and the desired signal can be calculated. The different processing steps are illustrated in detail in the following block diagram.

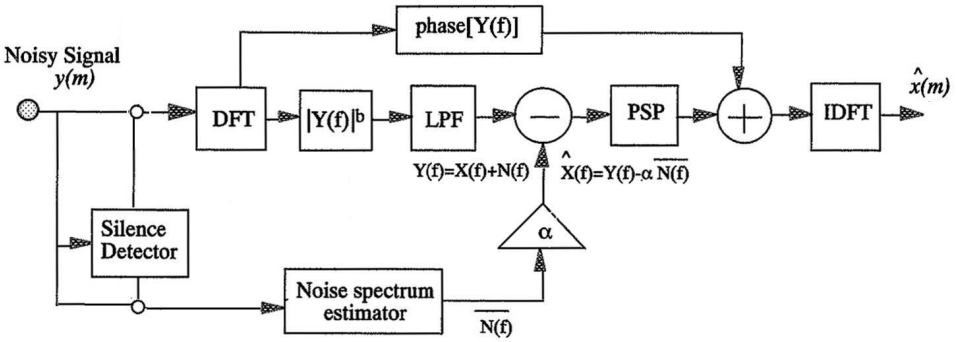

Figure 5: Spectral subtraction in detail

3 Gammatone Filterbank Subband Method

The second approach to the noise suppression problem uses a special filterbank decomposition on the input and operates in the time domain in sample-by-sample basis for each of the subband signals. Noise and speech detection is done separately in each subband, which allows more fine-grained isolation of the noise sections in the time-frequency plane in comparison to a global speech activity controlled estimation.

3.1 Gammatone filters

The filterbank used for the subband decomposition is a so called Gammatone filterbank or more specifically an Equivalent Rectangular Bandwidth (ERB) filterbank based on Gammatone filters. ERB is a filter width measure based on the psychoacoustic properties of hearing. Gammatone filters are low-order filters simulating the frequency response of the different sections of the cochlea in the inner ear. Gammatone filter decomposition results in a more useful frequency resolution when compared to a DFT approach, where the accuracy is too high in the high frequencies or too low in the low frequencies, depending on the overall DFT resolution. A more detailed description of Gammatone filterbanks can be found in the technical report by Slaney.

Figure 6 shows the magnitude responses of the gammatone filters of a 24-band ERB filterbank. The figure displays the high amount of crosstalk between the filters in comparison to traditional frequency selective filters. This corresponds to the behavior of the inner ear, where a relatively broad area of the cochlea is excited by a narrowband sound, but the excitation is most prominent at the section best tuned to the particular frequen-

cies of the input. Gammatone filters simulate approximately also the simultaneous (i.e. frequency-wise) masking phenomenon of the ear. When moving away from the filter center frequency, the attenuation increases only by the amount which is needed for the neighboring filters to mask the signal at their bands. This is advantageous in speech denoising as less denoising, and hence less added distortion, is applied for the suppression of audible noise. This minimization of the manipulation applied to the signal might be advantageous in reducing annoying artifacts of noise suppression.

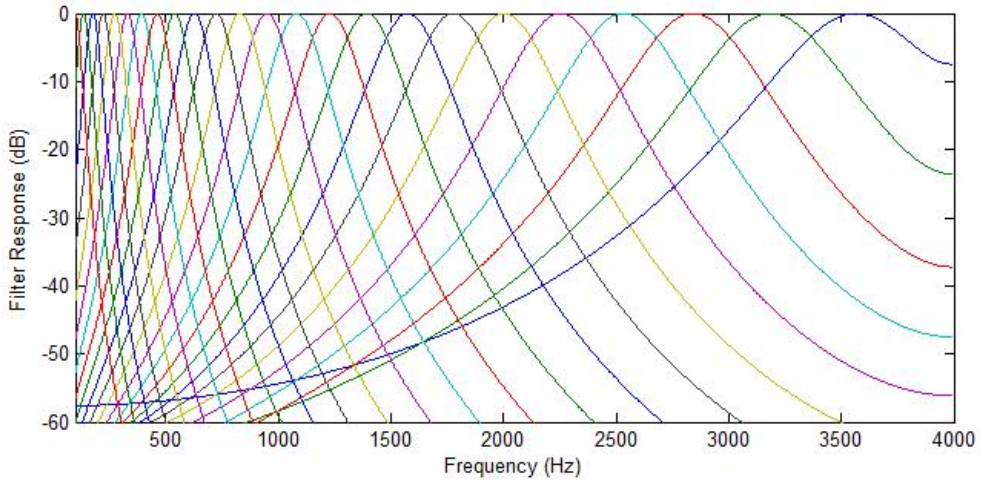

Figure 6: Magnitude response of a 24-band Gammatone filterbank.

3.2 Estimation of the noise

Before estimating the noise level in each subband, a power envelope estimation is applied by squaring and average-filtering the signal. Similar to the approach in section 2.1 a threshold level is set to each of the subband power envelope curves so that 70 percent of the envelope values stay under the threshold. This is based on the facts that the speech is assumed to have subband-wise larger instantaneous energy than the noise, and that the majority of a speech signal consists of pauses between syllables, words, and sentences. The estimated noise level is then calculated as described in section 2.1.

3.3 Generalized spectral subtraction

The noise attenuation is implemented with the so called generalized spectral subtraction method described in the paper by Virag. The method uses the a posteriori signal-to-noise ratio

$$SNR_{post} = \frac{|Y|^2}{|\hat{N}|^2} \tag{5}$$

with Y the noisy signal power and $\hat{N}$ the estimated noise power, to calculate a weighting function G for noise attenuation. The function gets values $0 \leqslant G \leqslant 1$ ranging from full attenuation to passing the unmodified input through, and is calculated as

$$G = \begin{cases} \left(1 - \alpha \cdot \left[\frac{|\hat{N}|}{|Y|}\right]^{\gamma_1}\right)^{\gamma_2}, \text{if } \left[\frac{|\hat{N}|}{|Y|}\right]^{\gamma_1} < \frac{1}{\alpha + \beta} \\ \left(\beta \cdot \left[\frac{|\hat{N}|}{|Y|}\right]^{\gamma_1}\right)^{\gamma_2}, \text{otherwise.} \end{cases} \quad (6)$$

α and β are the oversubtraction factor and the spectral floor factor as explained in sections 2.6.1 and 2.2, respectively. The proper choice of the parameters γ_1 and γ_2 allows choosing between different basic denoising methods as follows

$\gamma_1 = 1$ and $\gamma_2 = 1$: Magnitude subtraction
$\gamma_1 = 2$ and $\gamma_2 = 0.5$: Power spectral subtraction
$\gamma_1 = 2$ and $\gamma_2 = 1$: Wiener filtering

The method thus gives an easy way to choose between the spectral subtraction approaches and the Wiener filter, which is briefly described in the following section. A typical noisy signal and the resulting gain function averaged between the subbands are shown in Figure 7.

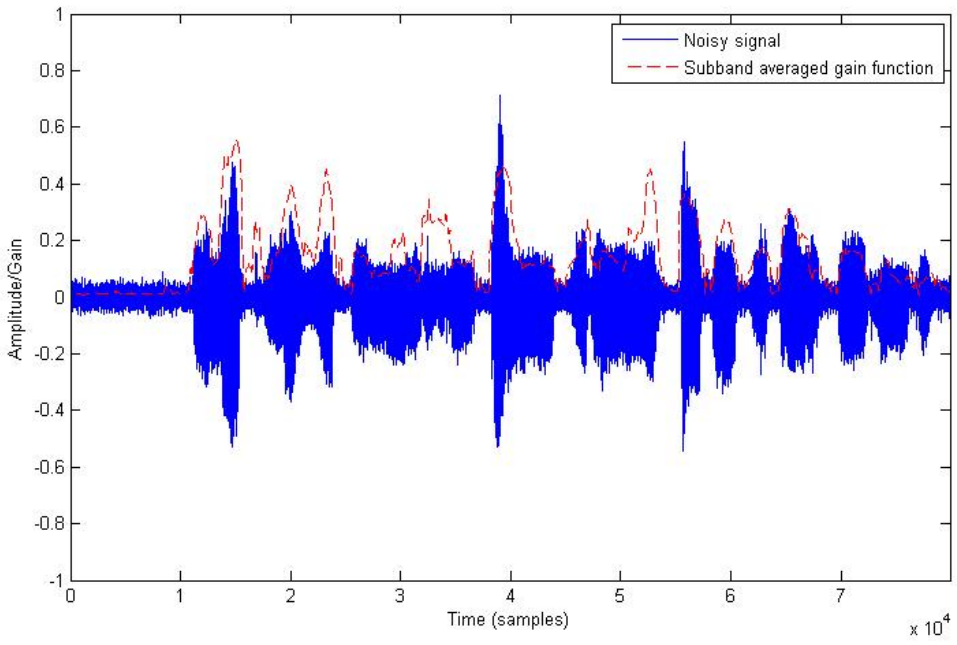

Figure 7: Noisy input signal and the resulting gain function.

3.3.1 Wiener filter

The Wiener filter is an adaptive error minimizing filter. Usually Wiener filters are implemented as FIR filters as the feedback of IIR filters combined with coefficient adaption would be a difficult combination to keep stable. With this being the case, Wiener filter output is just a linear convolution

$$\hat{x}(m) = \hat{\mathbf{h}}^T \mathbf{y}(m), \tag{7}$$

where $x(m)$ is the output sample at time index m, $\mathbf{h}$ the filter impulse response and $\mathbf{y}(m)$ input signal frame with the same length as the filter and the first element being the current input sample. An error signal is calculated by subtracting the filter output from the desired output

$$e(m) = x(m) - \hat{x}(m) = x(m) - \hat{\mathbf{h}}^T \mathbf{y}(m). \tag{8}$$

The filter coefficients are chosen to minimize a certain error based cost function, which can be mean square error (MSE), expectation of the absolute error, or expectation of some higher power of the absolute error. Choosing MSE as the optimization criteria results in a unique global minimum in the error surface.

$$J(\hat{\mathbf{h}}) = E\{e^2(m)\} \tag{9}$$

The gradient vector of the cost function with regard to the filter coefficients equals $-2E\{e(m)\mathbf{y}(m)\}$. The error mimimum is found by setting the gradient equal to a zero vector, which leads to the so called Wiener-Hopf equation

$$\mathbf{R}\hat{\mathbf{h}_\mathbf{o}} = \mathbf{p}, \tag{10}$$

where $\mathbf{R} = E\{\mathbf{y}(m)\mathbf{y}^T(m)\}$ is the correlation matrix of $\mathbf{y}(m)$ and $\mathbf{p} = E\{\mathbf{y}(m)x(m)\}$ the cross-correlation vector between $\mathbf{y}(m)$ and $x(m)$. Assuming nonsingularity of $\mathbf{R}$ the optimal filter weigths are

$$\hat{\mathbf{h}_\mathbf{o}} = \mathbf{R}^{-1}\mathbf{p}. \tag{11}$$

A more detailed explanation of the Wiener filter theory can be found in the book *Adaptive Filter Theory* by S. Haykin.

3.4 Summary

Figure 8 shows the block diagram of the implemented system. After denoising the subbands are simply summed up to form the output signal $\hat{x}(m)$. With the additional complexity of the subband method, the parameter optimization of the implemented system is more problematic than with the method described in section 2. The input noise suppression is somewhat better thanks to the noise estimation controlled individually in the subbands; there's no perceivable background noise increase during the speech segments. In addition to the parameters α, β, γ_1, γ_2, and noise threshold percentage, the averaging filter lengths of the input and noise estimate power envelopes seem to have a big effect on the system performance.

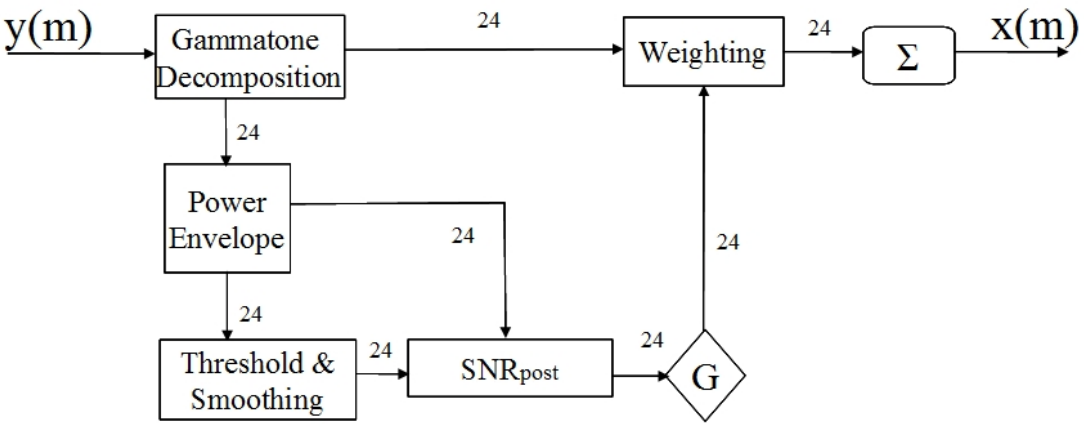

Figure 8: Block diagram of the Gammatone filterbank denoising system.

4 Sources

*1) S. V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction, Wiley and Teubner, 1996

*2) M. Lorber, Rauschverminderung zur Restauration von Audiosignalen, diploma thesis, IEM, 1997

*3) P. Vary, R. Martin, Digital Speech Transmission, Wiley, 2006

*4) N. Virag, Single Channel Speech Enhancement based on masking properties of the human auditory system, IEEE Vol.7, No.2, March 1999

*5) M. Slaney, An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank, Apple Computer Technical Report No. 35, Â© 1993, Apple Computer, Inc.

*6) S. F. Boll, Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Transactions on Acoustics, Speech, and Signal Srocessing, Vol. ASSP-27, No. 2, April 1979

*7) Chen et al., New Insights into the Noise Reduction Wiener Filter, IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.4, July 2006

*8) S. Haykin, Adaptive Filter Theory, 3rd Edition, Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1996

*9) http://cslu.cse.ogi.edu/nsel/wan_manuscript/spectsub.gif, February 25 2010