Speech Enhancement

for Mobile Communications

Harald Gustafsson

Ronneby, February 2000

Speech Enhancement

for Mobile Communications

Harald Gustafsson

Ronneby, February 2000

Department of Telecommunications and Signal processing
University of Karlskrona/Ronneby
S–372 25, Ronneby, Sweden

© Harald Gustafsson

Published 2000
Printed by Lund University, Reprocentralen
Lund 2000

Preface

This Licentiate thesis, *Speech Enhancement for Mobile Communications*, summarizes some of my research regarding speech enhancement. The work has been carried out at the University of Karlskrona/Ronneby in Ronneby and Ericsson Mobile Communications in Lund. This close relation to industry has been very fruitful. Three major applications of speech enhancement have been studied: noise reduction for mobile telephone handsfree accessories, noise reduction for handheld mobile telephones, and short delay noise reduction for frame-less telephone systems. The thesis consist of three parts:

H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction, Truly Linear Convolution and a Spectrum Dependent Adaptive Averaging Method", submitted to IEEE Transactions on Speech and Audio processing, June, 1999.

H. Gustafsson, S. Nordholm, I. Claesson, U. Lindgren, "Dual-Microphone Spectral Subtraction", Research Report 2000:2, University of Karlskrona/Ronneby, ISSN 1103–1581, February, 2000.

H. Gustafsson, S. Nordholm, I. Claesson, "Noise Suppression Employing Short Delay Processing", Proc. of ISPACS '99, Phuket, December 1999.

Parts of the papers have been presented as:

H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction Using Correct Convolution and a Spectrum Dependent Exponential Averaging Method", Research Report 15/98, University of Karlskrona/Ronneby, August 1998.

H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction Using Correct Convolution", Proc. of ICSV6, Copenhagen, July 1999.

H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction with Adaptive Averaging of the Gain Function", Proc. of EuroSpeech '99, Budapest, September 1999.

H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction Using Dual Microphones", Proc. of IWAENC '99, New York, September 1999.

The ideas presented have been filed for patent together with Ericsson as:

H. Gustafsson, S. Nordholm, I. Claesson, "Signal noise reduction by spectral subtraction with correct filtering", US Pat application 09/084,387.

H. Gustafsson, S. Nordholm, I. Claesson, "Signal noise reduction by spectral subtraction with exponential averaging", US Pat application 09/084,503.

H. Gustafsson, S. Nordholm, I. Claesson, "Dual microphone noise reduction using spectral subtraction", US Pat application 09/289,065.

H. Gustafsson, S. Nordholm, I. Claesson, "Signal noise reduction by time-domain spectral subtraction", US Pat application 09/289,555.

H. Gustafsson, S. Nordholm, I. Claesson, U. Lindgren, "Dual microphone noise reduction adaptive control", Patent pending.

Acknowledgements

I would like to thank many for helping and supporting me during my studies. I begin by thanking Professor Ingvar Claesson for being such an inspiring teacher. His teaching even during my undergraduate courses aroused my curiosity for signal processing. I would also like to thank Professor Sven Nordholm for being an excellent supervisor. They have given me a lot of support and inspiration together with constructive criticism to keep me on track. I am grateful to Björn Ekelund at Ericsson Mobile Communications, for agreeing to have me as an Industrial PhD student at Ericsson and HK/R. I would like to thank my managers, Richard Elofsson, Jim Rasmusson and Dr Ulf Lindgren at Ericsson for supporting me during my studies and work. Special warm thanks go to Dr Ulf Lindgren for always having time for discussions and research.

Further, I thank my colleagues at the Department of Telecommunications and Signal Processing at HK/R and Ericsson Mobile Communications in Lund.

I would like to thank my family for their continuous support and interest. Last but definitely not least I would like to thank Anja, my wife, for all her love and tremendous support.

Harald Gustafsson
Ronneby, February 2000

Contents

Introduction 1–4

Part A: *Spectral subtraction, truly linear convolution and a spectrum dependent adaptive averaging method* 5–34

Part B: *Dual-microphone spectral subtraction* 35–66

Part C: *Noise suppression employing short delay processing* 67–78

Introduction

During conversation, both hearing and speaking adapt to the background noise in a noisy environment. It is therefore possible to have a conversation in quite disturbing background noise environments. However, when the conversation takes place over the telephone disturbances are more annoying. The disturbances are a problem since the brain will not get the extra visual and other background information when interpreting the speech. The speech signal transmitted to the other party is picked up by a microphone connected to the telephone. The microphone signal contains both *speech* and *noise* at some *ratio* (Speech to Noise Ratio, SNR) depending on, for example, how far the microphone is mounted from the speaker's mouth.

In vehicles, it is common to use a telephone handsfree accessory. The main motivation for this is to facilitate both hands on the steering wheel when driving. The drawback is that the telephone microphone is situated at a farther distance (50 cm) from the speaker's mouth as compared to handheld telephony, in a high noise level environment. To increase the SNR and to allow for the listener to grasp the speech clearly, a speech enhancement method should be applied.

The basis for all the methods presented in this thesis is a novel spectral subtraction algorithm. Spectral subtraction is a speech enhancement method. The reduction of noise is accomplished by subtracting an estimate of the power spectrum of the noise from the power spectrum of the input signal while maintaining the noisy speech signal phase information. This can also be viewed as a filtering operation. The filter is shaped so that frequencies with high SNR may pass, but frequencies with low SNR are proportionally reduced. This results in a speech enhanced signal with a higher overall SNR. The novel method improves the conventional spectral subtraction method by exchanging the non-causal circular convolution with a purely linear causal filtering.

The filter used in spectral subtraction is applied in the frequency-domain and changes over time for each segment of the microphone signal. The signal segment is transformed to the frequency-domain by a Fast Fourier Transform (FFT) resulting in a frequency-domain segment. Conventionally the filter is of the same order as the signal segment and the frequency-domain segment, yielding a circular convolution. Since the segments and the filter are of the same length, causal properties cannot be imposed on the filter. By using a shorter filter length, the filter can be estimated with less variance. A shorter

filter length also has the benefit of providing a purely linear filtering possibility in the frequency-domain when the added length of the signal segment and the filter is shorter than the FFT length. The purely linear filtering does not introduce sharp edges between segments as with conventional spectral subtraction. These discontinuities can sometimes be heard as click sounds at segment boundaries. In conventional spectral subtraction these click sounds can be reduced by introducing a time-window, which gradually mutes the ends of a signal segment, in combination with an overlapping between segments. The introduced overlap generally results in a delay of the processed signal of approximately 10 milliseconds.

With the proposed approach causal properties can be imposed on the filter. When the causality is imposed as a minimum phase, a delay of less than approximately half a millisecond is introduced. In part A of this thesis, the modified spectral subtraction method is presented in detail.

Part A also contains a new method for reducing fluctuations of the gain filter used in spectral subtraction. The amplitude frequency response of the filter, *the gain function*, can be smoothed over segments only when the frequency segments are similar. The proposed method applies an adaptive exponential averaging on the gain function, controlled by the normalized spectral discrepancy between the current amplitude frequency segment and the estimated amplitude frequency segment of the background noise signal. When the discrepancy is small, for example during noise only periods, the averaging has a long time span exploiting a large number of segments for estimation of the gain function. The adaptive control enables a reduced averaging time span during high-energy speech, thereby preventing a smearing of the speech signal.

In part B a new method for noise reduction in handheld mobile telephones is presented. When mobile phones are designed smaller, the microphone become situated farther from the speaker's mouth, thereby causing a decrease in the microphone signal SNR. A substantial decrease in SNR must be compensated with speech enhancement methods in order to maintain a comfortable speech quality.

In the vehicle handsfree approach described in part A it is possible to estimate the noise frequency segments during speech pauses, since the background generally has those same properties during speech periods. In the handheld case the user may be situated in environments that are not even approximately similar during speech and silent periods, since the properties may change quickly. The proposed handheld scheme makes use of a secondary microphone even farther away from the speaker's mouth, continuously estimating the background.

The algorithm consists of two pre-processing and one final processing

spectral subtraction algorithm. Although spectral subtraction normally is described as a speech enhancement algorithm, it can also be exploited for just the opposite, as a noise enhancement algorithm. The pre-processing spectral subtraction algorithms are employed to enhance the speech signal amplitude spectrum from the primary microphone and the noise signal amplitude spectrum from the secondary microphone. This is accomplished by subtracting the refined estimate from the other pre-processing spectral subtraction algorithm in a coupled way. The refined noise estimate is used in the final spectral subtraction algorithm to calculate the speech enhancement filter. The final spectral subtraction algorithm filters the primary microphone signal, producing the output signal. The output speech signal is *not* taken from the pre-processor since the refined speech signal of the pre-processor is optimized in order to have a reduced noise level, thereby sacrificing speech quality.

Finally, in part C frame-less systems are considered. Systems that work in real-time bidirectionally, such as telephone conversations, impose higher demands on the delay introduced by the system. Telephone systems typically introduce a delay between 10 ms to 100 ms. Digital telephone systems, for example the GSM mobile phone system, work on segments, or *frames*, of the microphone signal. The methods described in part A and B, are based on the presumption that the speech signal can be processed frame-wise without introducing an extra stacking delay. For analog telephone systems, no segmentation is performed. If the methods of part A and B were used for analog systems the microphone signal would have to be segmented, thereby introducing an extra delay. Such a delay would not comply with the specifications for most analog telephone systems. In part C of the latter thesis a solution to this problem is presented, where the general idea of the methods in part A and B can be used in analog systems as well. The analysis of the signal and the calculation of the shape of the filter can both be used on a frame-wise basis without introducing a delay. It is the actual filtering of the microphone signal that introduces the delay of the signal. When this filtering is instead performed in the time-domain, an extra stacking delay can be avoided. Further, if a minimum phase design is used to construct the filter, as previously suggested, the filtering delay will be less than half a millisecond also for analog telephone systems. The drawback with this solution is that the filter calculated for the present frame must be applied to the following frame. The result of part C shows that the filter changes at such slow rate that it is possible to use the filter from the previous frame with only a small degradation of the speech enhancement performance.

Part A

Spectral subtraction, truly linear convolution and a spectrum dependent adaptive averaging method

This part is published as: H. Gustafsson, S. Nordholm, I. Claesson, "Spectral Subtraction, Truly Linear Convolution and a Spectrum Dependent Adaptive Averaging Method", submitted to IEEE Transactions on Speech and Audio processing, June, 1999.

Spectral subtraction, truly linear convolution and a spectrum dependent adaptive averaging method

Harald Gustafsson, Sven Nordholm and Ingvar Claesson

Abstract

In handsfree speech communication the signal to noise ratio is often poor, which makes it difficult to have a relaxed conversation. By using noise suppression, the conversation quality can be improved. This paper describes a noise suppression algorithm based on spectral subtraction. The method employs a noise and speech dependent gain function for each frequency component. Proper measures have been taken to obtain a corresponding causal filter and also to ensure that the circular convolution originating from FFT filtering yields a truly linear filtering. A novel method that uses spectrum-dependent adaptive averaging to decrease the variance of the gain function is also presented. The result obtained is a 10 dB background noise reduction, with improvement in speech quality and reduction of noise artifacts compared with conventional spectral subtraction methods.

1 Introduction

Today, mobile phones are commonly used in our society. People tend to use phones in all environments, with vehicles being no exception. Traffic safety authorities encourage people to use handsfree accessories in order to have both hands available and focus better on the driving. When using handsfree accessories, the microphone is situated as far as one meter from the speaker, so that signal processing methods to suppress the background noise and the acoustic echo for the far-end speaker become necessary. This paper is directed towards the noise reduction problem and relies on ideas from the spectral subtraction algorithm [1].

Spectral subtraction employs estimates of the noise spectrum and the noisy speech spectrum to form an SNR(f)-based gain function. The gain function is multiplied by the input spectrum and will suppress frequency components with low SNR(f). The main disadvantage using conventional spectral subtraction algorithms is the resulting "musical tones" which disturb not only the listener but also speech coding algorithms. To solve this problem spectral smoothing has been suggested, resulting in reduced variance and resolution. Other suggested improvements of conventional spectral subtraction are to use the Bartlett method on fairly long sequences of samples to reduce the variance of the spectrum [2], or to hide the musical tones phenomenon using the masking properties of the auditory system [3], [4]. Another known method to reduce the musical tones is to use an over-subtraction factor (larger than one) in combination with a spectral floor [5]. This method has the disadvantage of degrading the speech when musical tones are sufficiently reduced.

The presently suggested method deals with 4 different issues:

- The variability of the gain function.
- Block effects from the FFT- and IFFT-operations.
- Non-causal filtering due to the zero-phase gain function.
- Short delay processing.

The musical tones are mainly due to variance in the spectrum estimates, [6]. The variance of the gain function is reduced by dividing the current input block into sub-blocks and performing a lower resolution spectrum estimate. This process results in a gain function that has lower variance and also lower resolution. The averaging is performed on both the noisy speech signal and the background noise. The noise estimate is averaged over several blocks

where only noise is present while the noisy speech is averaged only over the current block.

To further reduce the variability of the gain function, an adaptive exponential averaging has been introduced. The adaptive exponential averaging of the gain function employs a discrepancy measure between the noisy speech spectrum estimate and the noise spectrum estimate. When the input signal energy level is stationary, the discrepancy is low, thus the gain function may be substantially averaged resulting in low artificial noise. A high energy speech spectrum, i.e. when the discrepancy is large, will result in a reduced averaging, and in the extreme no averaging at all. Even though the averaging is reduced, the high energy speech will mask the artificial noise. Good sound quality will still be perceived, indicating that a gain function with low variability is obtained when it is needed. The proposed method removes the musical tones to such a degree that it is possible to use under-subtraction of the noise spectrum which in turn results in very low speech degradation.

Since the spectral subtraction techniques are block based and use FFT:s, the block effects must be considered. The FFT corresponds to a critical sampled filter bank, see [7]. There will be discontinuities between blocks which are conventionally masked by windowing the input samples using a Hanning window and introducing an overlap. This overlap results in a delay often of as much as 50 percent of the frame length. The discontinuities have their origin in the circular convolution (which comes from the FFT and IFFT operations) and the gain function which corresponds to a non-causal zero phase filter of the same length as the FFT. With this observation in mind, a proper block length of the FFT is suggested to obtain a correct linear convolution combined with an interpolated low resolution gain function. By imposing a linear or minimum phase constraint on the gain function corresponding to a causal filter. The conventional gain function in spectral subtraction, which is only an amplitude function, corresponds to a non-causal zero-phase filter. The benefits of the proposed method over the windowing method is that it avoids the discontinuities and the delay is reduced to merely a few samples.

The suggested methods are presented in Section 2. This section also includes a short introduction to conventional spectral subtraction. Further in Section 3 the experimental results are presented. Section 4 concludes the paper.

2 Spectral Subtraction

Spectral subtraction relies upon the assumption that the background noise signal has an almost constant magnitude spectrum and the speech signal

is short-time stationary. Furthermore, the background noise is considered additive and uncorrelated to the speech signal. Let $s(n)$, $w(n)$ and $x(n)$ represent the speech signal, noise signal and noisy speech signal, respectively,

$$x(n) = s(n) + w(n). \tag{1}$$

The short-time power spectral density relation is thus

$$R_x(f,i) = R_s(f,i) + R_w(f,i) \tag{2}$$

where $f \in [0, N-1]$ is a discrete variable enumerating the frequency bins and i is a time block index. The short-time spectral density is estimated by using the periodogram

$$\hat{R}_{x,N}(f,i) = \frac{1}{N} \left| \mathcal{F}\{\mathbf{x}_N(i)\} \right|^2 \tag{3}$$

where $\mathbf{x}_N(i)$ is a vector containing the i:th block of N data samples and $\mathcal{F}$ is the FFT operation. For convenience, the magnitude spectrum estimate is defined as $\hat{P}_{x,N}(f,i) = \sqrt{\hat{R}_{x,N}(f,i)}$. The background noise magnitude spectrum can be estimated over a longer time frame by

$$\bar{P}_{w,N}(f,i) = \begin{cases} \mu \bar{P}_{w,N}(f,i-1) + (1-\mu)\hat{P}_{x,N}(f,i) & \text{, noise only} \\ \bar{P}_{w,N}(f,i-1) & \text{, speech and noise} \end{cases} \tag{4}$$

where μ is the exponential averaging time constant. The speech pauses are detected by a Voice Activity Detector, VAD.

The spectral subtraction operation corresponds to a time varying filtering operation

$$Y_N(f,i) = G_N(f,i) X_N(f,i) \tag{5}$$

over a block of samples where capital letters denote the FFT:s of the corresponding time blocks. The multiplication of the input signal and the gain function corresponds to a circular convolution in the time-domain. Since $G_N(f,i)$ corresponds to a non-causal filter it should be phase shifted to obtain causality and to avoid circular effects. The effect of circular convolution is aliasing in the time-domain, and the non-causal filtering leads to discontinuities between blocks, thereby giving rise to inferior speech quality. The gain function is given by

$$G_N(f,i) = \left(1 - k \cdot \frac{\bar{P}^a_{w,N}(f,i)}{\hat{P}^a_{x,N}(f,i)} \right)^{\frac{1}{a}} \tag{6}$$

where k is the subtraction factor and a controls whether magnitude or power spectral subtraction is used.

2.1 Truly Linear Convolution

The time-domain aliasing problem inherited from periodic circular convolution can be solved by letting the gain function $G_N(f,i)$ originate from a shorter time function and leaving time space for filtering of the input signal $X_N(f,i)$ by padding with zeros. To construct the spectrum of order L, an input signal frame, $\mathbf{x}_L(i)$, of duration $L < N$, is used. By zero padding the frame, $\mathbf{x}_L(i)$, to the full block length N, the spectrum, $X_{L\uparrow N}(f,i)$, is constructed.

The gain function in Equation (6) contains a division. Therefore, an averaged lower resolution gain function, $G_M(f,i)$, should be used in order to avoid extreme variability (close to zero in the denominator). Both the averaged background noise spectrum estimate, $\bar{P}_{w,M}(f,i)$, and the recent frame noisy speech spectrum estimate, $\hat{P}_{x,M}(f,i)$, employed in the computation of the gain function have block length $M \ll N$

$$G_M(f,i) = \left(1 - k \cdot \frac{\bar{P}^a_{w,M}(f,i)}{\hat{P}^a_{x,M}(f,i)}\right)^{\frac{1}{a}}. \tag{7}$$

The shorter periodogram estimates are computed by using sub-blocks of the input frame, $\mathbf{x}_L(i)$, combining a Bartlett method [8], which is used to decrease the variance of the estimated spectrum, and the rule according to Equation (4) in order to further reduce the noise artifacts. This averaged gain function, $G_M(f,i)$, has a corresponding impulse response, $g_M(n,i)$, of length $M < N - L$, which gives space to avoid circular effects. In order to reconstruct a gain function that matches the number of FFT bins, N, the gain function is interpolated from the shorter gain function, $G_M(f,i)$, to form $G_{M\uparrow N}(f,i)$. Although $G_{M\uparrow N}(f,i)$ has N frequency bins, the corresponding impulse response is only of length M. We utilize the lower resolution of $G_{M\uparrow N}(f,i)$ to accomplish a truly linear filtering. The resulting output is obtained by using overlap-add and an inverse FFT of

$$Y_N(f,i) = G_{M\uparrow N}(f,i) X_{L\uparrow N}(f,i). \tag{8}$$

The pure amplitude gain function, $G_{M\uparrow N}(f,i)$, is a real function, thus non-causal and zero-phase. Causality is therefore imposed to obtain truly linear filtering. Causality is constructed in two ways, either with linear phase or minimum phase. The causal linear phase filter corresponding to an amplitude function [8], is achieved by imposing a linear phase on the gain function.

Alternatively, the relationship between the amplitude function and phase function for a minimum phase filter can be utilized. The causal minimum

phase filter is obtained from the gain function by using a Hilbert transform relation [9]. The Hilbert transform relation implies a unique relationship between the real and imaginary parts of a complex function. It can also be utilized for a relationship between magnitude and phase when the logarithm of the complex gain function is used. For discrete-time systems the minimum phase properties can in general only be fulfilled approximately.

2.2 Further Variance Reduction of the Gain Function

The reduced resolution has been used to reduce the variations of the spectrum estimates which form the gain function, $G_M(f,i)$. The variations may be further decreased by using an adaptive exponential averaging of the gain function,

$$\bar{G}_{M,1}(f,i) = \alpha_1(i) \cdot \bar{G}_{M,1}(f,i-1) + (1-\alpha_1(i)) \cdot G_M(f,i) \tag{9}$$

where $\alpha_1(i)$ is an adaptive averaging time parameter derived from a spectral discrepancy measure and $\bar{G}_{M,1}(f,i)$ is the adaptively averaged gain function. How this adaptive gain function calculation enters into the total scheme is illustrated by Figure 1(b). The adaptive averaging time parameter, $\alpha_1(i)$, is derived from a spectral discrepancy measure, $\beta(i)$, where

$$\alpha_1(i) = 1 - \beta(i). \tag{10}$$

The exponential averaging time in frames is approximately

$$\tau_1(i) \approx \frac{1}{1-\alpha_1(i)}. \tag{11}$$

The spectral discrepancy measure, $\beta(i)$, depends on the relation between the current block spectrum, $\hat{P}_{x,M}(f,i)$, and the current averaged noise spectrum, $\bar{P}_{w,M}(f,i)$

$$\beta(i) = \min \left\{ \begin{array}{c} \dfrac{\sum_{f=0}^{M-1} \left| \hat{P}_{x,M}(f,i) - \bar{P}_{w,M}(f,i) \right|}{\sum_{f=0}^{M-1} \bar{P}_{w,M}(f,i)}, \quad 1 \end{array} \right\}. \tag{12}$$

A small spectral discrepancy yields a large $\alpha_1(i)$ and thus a longer averaging time of the gain function, $G_M(f,i)$. This corresponds to a stationary background noise situation. A large spectral discrepancy should result in a shorter averaging time, or no averaging of the gain function, $G_M(f,i)$. This corresponds to situations where speech or highly varying background noise is present.

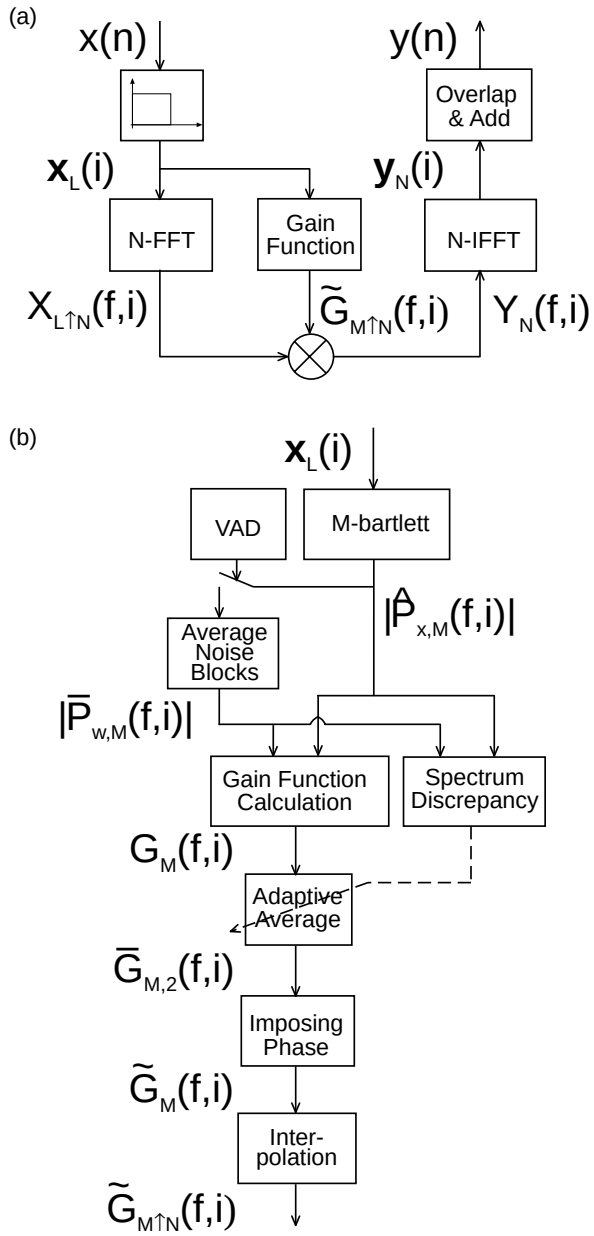

Figure 1: (a) Outline of the improved spectral subtraction algorithm. (b) A detailed view of the gain function calculation in (a), consisting of the parts facilitating the new causal truly linear filtering and adaptive exponential averaging.

Precaution must be taken when the input signal goes from a noisy speech period to a background noise period, where the averaging time parameter, $\alpha_1(i)$, increases so that the effective averaging time of the gain function, $\bar{G}_{M,1}(f,i)$, also increases. However, a direct increase of the averaging time would result in an audible "shadow voice," since the gain function suited for a speech spectrum would linger for a long period. Accordingly, the averaging time should only be allowed to increase slowly, allowing the gain function to adapt to the stationary input. Thus, by inserting a conditional averaging of the adaptive averaging time constant, $\alpha_2(i)$, of the gain function, this effect can be achieved,

$$\begin{aligned} \text{if} \quad & (\alpha_2(i-1) - \alpha_1(i)) < 0 \ \text{ then} \\ & \alpha_2(i) = \gamma_c \alpha_2(i-1) + (1-\gamma_c)\alpha_1(i) \\ \text{else} \quad & \\ & \alpha_2(i) = \alpha_1(i) \end{aligned} \tag{13}$$

Thus, as long as the averaged time constant, $\alpha_2(i)$, has not reached the level of the first time constant, $\alpha_1(i)$, it will increase slowly according to γ_c. In other situations $\alpha_2(i)$ follows the first time constant, $\alpha_1(i)$, directly.

The variance reduced gain function is thus given by

$$\bar{G}_{M,2}(f,i) = \alpha_2(i) \cdot \bar{G}_{M,2}(f,i-1) + (1-\alpha_2(i)) \cdot G_M(f,i) \tag{14}$$

having an approximate effective averaging time

$$\tau_2(i) \approx \frac{1}{1-\alpha_2(i)}. \tag{15}$$

Concluding, the adaptive averaging time constant, $\alpha_2(i)$, can decrease rapidly but increases only slowly, having the effective averaging time,

$$\tau_{\alpha_2}(i) \approx \frac{1}{1-\alpha_2(i)}. \tag{16}$$

This averaging scheme forces a slow increase of the averaging time, $\tau_2(i)$, and therefore the gain function can quickly adapt to the new input signal reducing the "shadow voices".

2.3 Operation of the Adaptive Gain Function

In this section the adaptive averaging of $G_M(f,i)$ is illustrated for different input signal scenarios. During background noise periods the gain function can

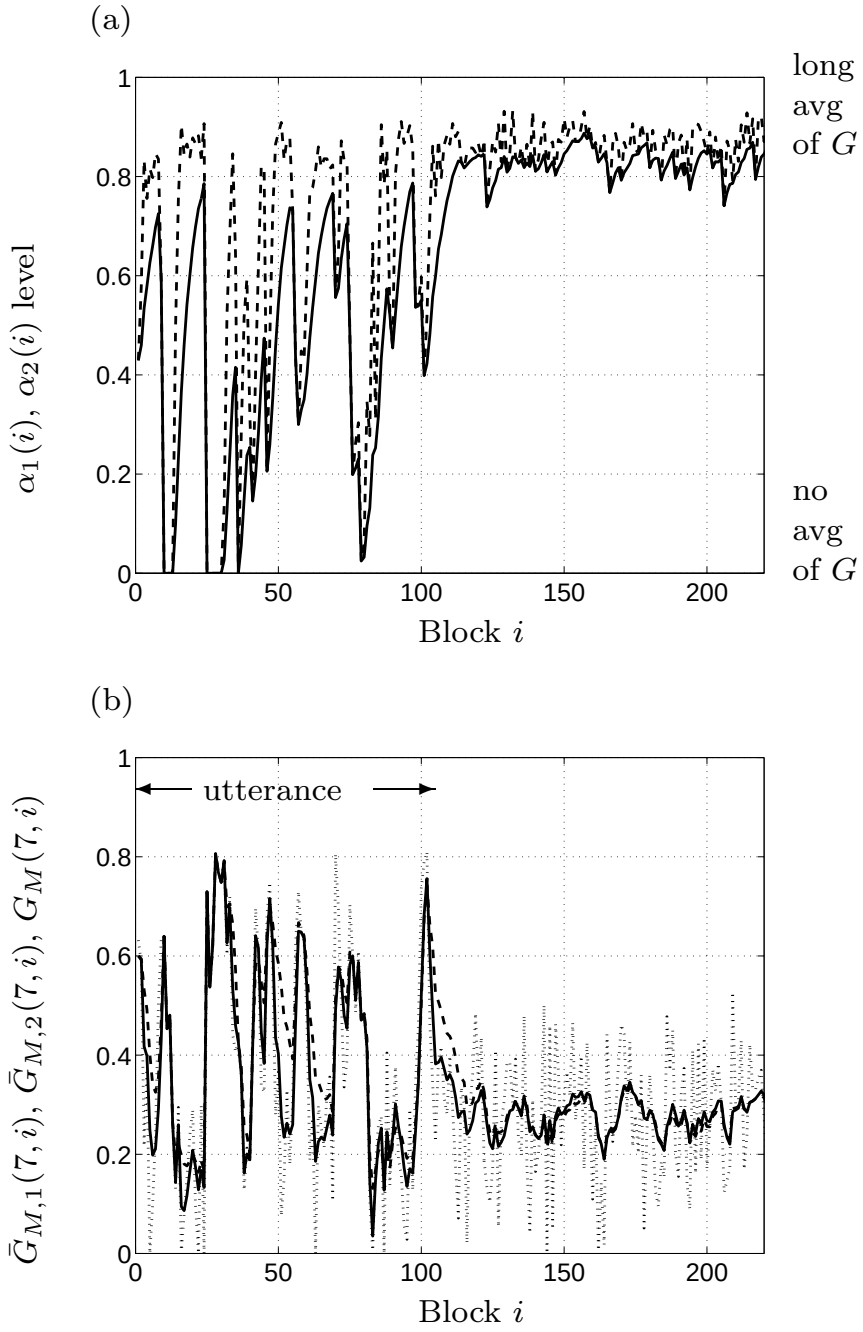

Figure 2: (a) Averaging time parameter vs. block i. Solid line: $\alpha_2(i)$; dashed line: $\alpha_1(i)$. (b) Frequency bin 7 of the gain function vs. block i. Solid line: $\bar{G}_{M,2}(7,i)$; dashed line: $\bar{G}_{M,1}(7,i)$; dotted line: using zero averaging of the gain function $G_M(7,i)$. The parameter $\gamma_c = 0.8$.

be averaged over a long time period which decreases the variance, provided the background noise spectrum has a steady amplitude spectrum. A long averaging time, $\tau_2(i)$, corresponds to a high level of $\alpha_2(i)$ as can be seen for the last 100 blocks in Figure 2. Signal amplitude level changes give rise to a decreased averaging time parameter, $\alpha_2(i)$. The adaptive averaging method will thus decrease the effective gain function averaging time, $\tau_2(i)$, until the signal amplitude spectrum level has settled at a new level. This behavior enables prompt response to amplitude spectrum level changes and gives a decrease in variance during stationary input signal periods. When the signal amplitude spectrum varies rapidly and exhibits time-varying spectral peaks the input signal is usually a speech signal. Consequently, adjacent gain functions have different shapes. Thus, the adaptive averaging time should be kept to a minimum during these periods. This effect can be observed in Figure 2, e.g. for blocks 75-85 by the similarity between the gain functions $\bar{G}_{M,1}(f,i)$ and $\bar{G}_{M,2}(f,i)$ with averaging and $G_M(f,i)$ without averaging. The short averaging time, $\tau_2(i)$, is due to the low levels seen in α_2, e.g. for blocks 75-85 in Figure 2. At the end of a spectral peak, e.g. at block 100, the parameters $\alpha_1(i)$ and $\alpha_2(i)$, as well as, $\bar{G}_{M,1}(f,i)$ and $\bar{G}_{M,2}(f,i)$ differs. As can be seen in Figure 2, $\bar{G}_{M,2}(f,i)$ decreases faster than $\bar{G}_{M,1}(f,i)$ after a spectral peak and adapts to the new input signal situation more quickly. This gives a reduction of "shadow voices" which occurred in the first averaged scheme using $\bar{G}_{M,1}(f,i)$. Remember that $\alpha_2(i)$ increases more slowly than $\alpha_1(i)$ and a higher averaging time parameter value gives a longer averaging time in the averaging of $G_M(f,i)$.

3 Results

The results show improvements in the speech quality and residual background noise quality compared with other spectral subtraction approaches, while still achieving similar or improved reduction in noise level. Another benefit is the short delay.

The voice activity detector, VAD, is a vital part of the noise reduction method. In these experiments the GSM VAD [10] has been used to detect when speech is present in the input signal. The signals used in this section were created using separate recordings of the speech signal and the background noise signal. The speech recordings were performed in a quiet stationary car while car noise was injected through the speaker's headphones in order to take the Lombard effect into consideration. The noise sequences were recorded using the same equipment in a car driving on a highway at the speed of 110 km/h. Both the noise and speech signals were sampled at 8 kHz and

filtered through a telephone bandwidth filter. The inputs and results are presented as sound files at our web-site: `http://www.its.hk-r.se/research`.

3.1 Parameter Choices

The parameter choices derived in this section are mainly directed towards a handsfree GSM mobile phone solution for vehicle use.

First the frame length, L, is chosen. In this case L is already fixed from the GSM mobile phone system specification, $L = 160$, which gives 20 ms frames. The next parameter to be determined is the length, M, of the periodogram. In order to reduce variance, M should be small. Since an FFT is used to compute the periodograms, the length M should be a power of two. Experiments has determined that $M = 32$ is suitable. This yields a length $L + M = 160 + 32 = 192$, which should be less than N, the FFT length. Thus, $N = 256$ which is the nearest power of two.

Magnitude spectral subtraction, when $a = 1$, is chosen which gives strong noise reduction and still good sound quality. The subtraction factor is chosen as $k = 0.7$, i.e. under-subtraction which will preserve the good sound quality.

The time constant, γ_c, controls the exponential decaying of the averaged spectral discrepancy, $\overline{\beta}(i)$. Experiments indicates that $\gamma_c = 0.8$ is a good choice for preventing "shadow voices".

3.2 Degree of Noise Reduction

There is a trade off between the noise suppression and speech quality. By choosing more radical parameters values a and k, further improvement of the noise reduction can be achieved at the expense of speech quality. Typical speech signals and background noise signals are shown in Figures 3 and 4, respectively. The combined input signal is presented in Figure 5. The noise reduced output is illustrated in Figure 6.

Since the gain function is a linear filter, although time-varying between blocks, the speech signal and background noise signal can be filtered separately for evaluation purposes. We then use a pre-calculated gain function where the gain function calculations have been performed on the combined input data. The sum of the outputs will result in the same output signal as if the noisy speech were filtered

$$\begin{aligned} y(n) = g(x(n), n) = g(s(n) + w(n), n) = \\ g(s(n), n) + g(w(n), n) = y_s(n) + y_w(n) \end{aligned} \tag{17}$$

where $y_s(n)$ and $y_w(n)$ are the processed speech signal and background noise signal, respectively, and $g(\bullet, n)$ denotes a linear time-varying operator. It

is possible now to calculate the block energy for the input speech signal, $p_s(i)$, input background noise signal, $p_w(i)$, processed speech signal, $p_{y_s}(i)$, and processed background noise signal, $p_{y_w}(i)$. To evaluate performance, block-wise SNR and noise reduction measures are defined as

$$\Delta_{\mathrm{SNR}}(i) = \frac{p_w(i)}{p_{y_w}(i)} \cdot \frac{p_{y_s}(i)}{p_s(i)} = \frac{\mathrm{SNR}_{out}(i)}{\mathrm{SNR}_{in}(i)} = \mathrm{SNR}_{out}(i)[\mathrm{dB}] - \mathrm{SNR}_{in}(i)[\mathrm{dB}] \quad (18)$$

which is the signal to noise ratio improvement for block i. The signal to noise ratio before the spectral subtraction is denoted by $\mathrm{SNR}_{in}(i)$ and the signal to noise ratio after processing is denoted by $\mathrm{SNR}_{out}(i)$,

$$\mathrm{SNR}_{in}(i) = \frac{p_s(i)}{p_w(i)}, \quad (19)$$

$$\mathrm{SNR}_{out}(i) = \frac{p_{y_s}(i)}{p_{y_w}(i)}. \quad (20)$$

Finally, the Noise Reduction, $\mathrm{NR}(i)$, is defined as

$$\mathrm{NR}(i) = \frac{p_w(i)}{p_{y_w}(i)}. \quad (21)$$

The SNR measures are only valid during speech periods. A histogram of the percentage of blocks with a certain noise reduction during noise only periods is presented in Figure 7. As seen the noise reduction during noise periods are mostly in the range of 10–12 dB. During speech periods the achieved noise suppression is dependent on $\mathrm{SNR}_{in}(i)$. Figure 8 presents a 2-dimensional histogram of the percentage of blocks with a certain $\mathrm{SNR}_{in}(i)$ achieving a certain noise reduction during speech periods. For blocks with higher $\mathrm{SNR}_{in}(i)$ it is more difficult to achieve a noise reduction since the gain function will be close to 1. Noise reduction is, however, not as crucial in periods where the speech signal clearly dominates, due to masking effects. In Figure 9, a 2-dimensional histogram of the percentage of blocks achieving a certain SNR improvement, $\Delta_{\mathrm{SNR}}(i)$, with a certain $\mathrm{SNR}_{in}(i)$ during speech periods is presented. The $\Delta_{\mathrm{SNR}}(i)$ is in the vicinity of 3 dB during speech periods. The noise suppression during noise periods is more than 10 dB.

3.3 Avoiding Undesirable Periodicity Effects

The experimental results presented in this section show the importance of having the appropriate impulse response length of the gain function as well as causal properties.

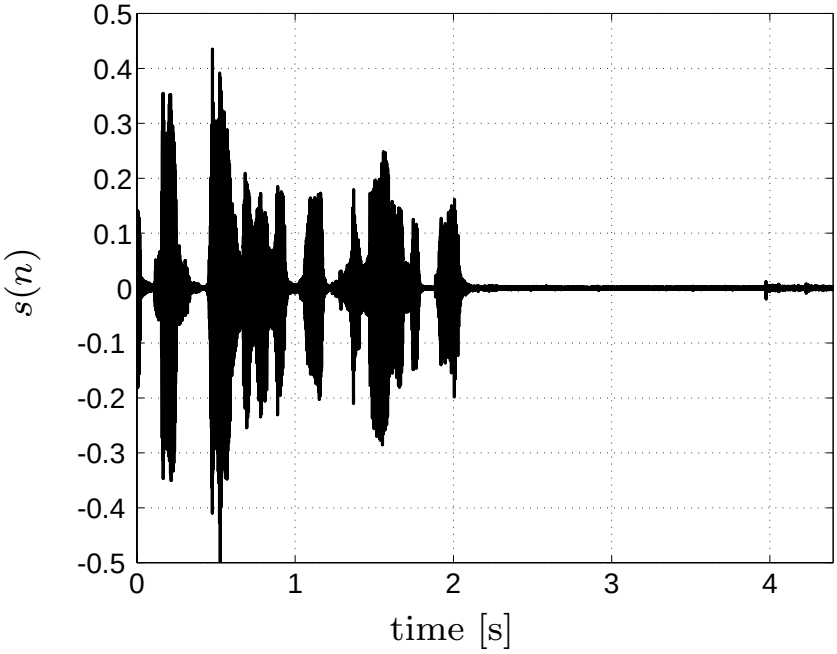

Figure 3: Input speech signal.

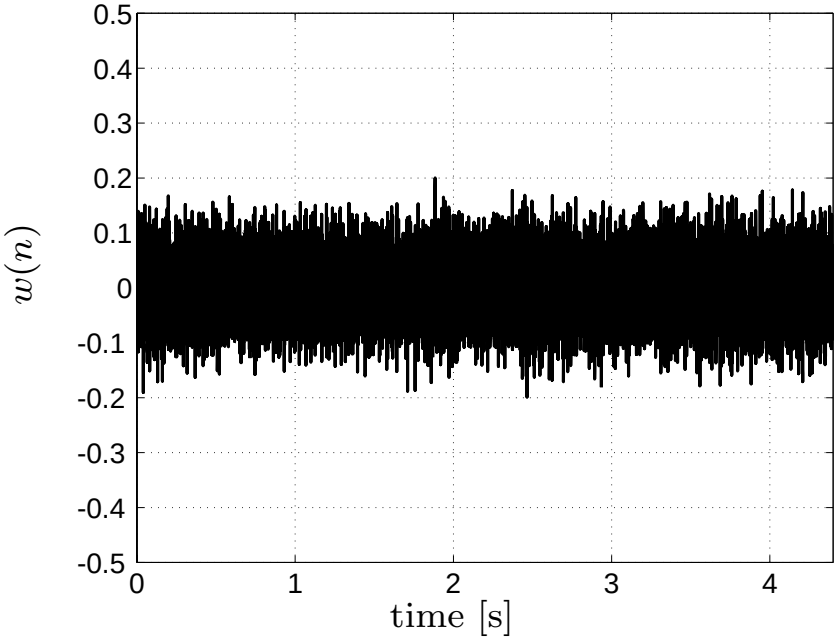

Figure 4: Input background noise signal.

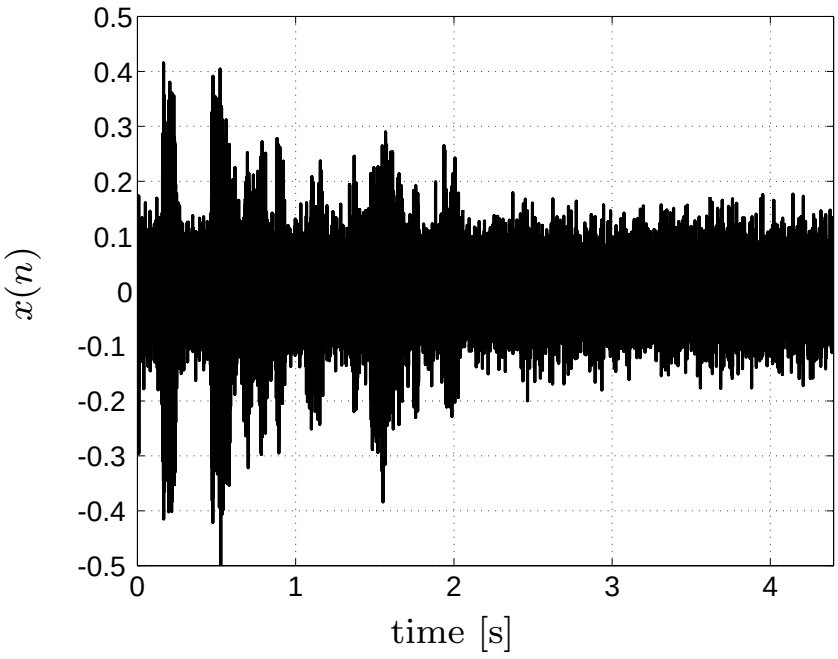

Figure 5: Input noisy speech signal.

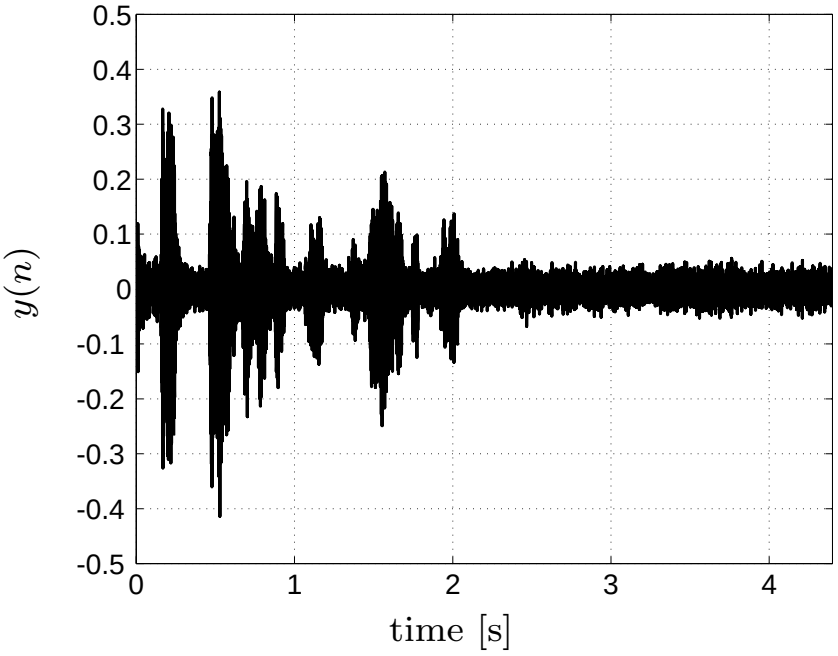

Figure 6: Output speech when noise reduction is employed.

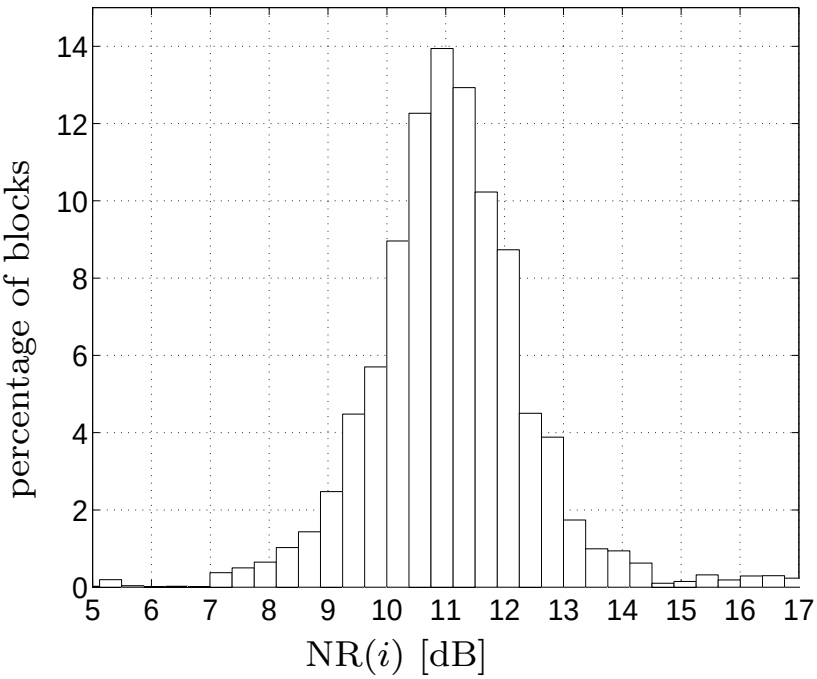

Figure 7: Histogram over percentage of blocks with a certain noise reduction during noise only periods.

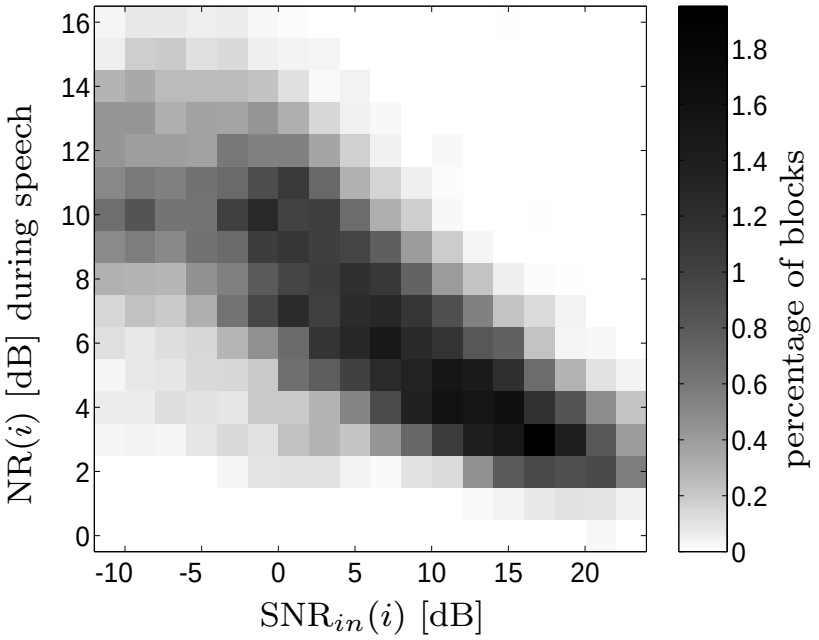

Figure 8: 2D-histogram over percentage of blocks with certain noise reduction, SNR$_{in}(i)$, and achieving a certain NR(i) during speech periods.

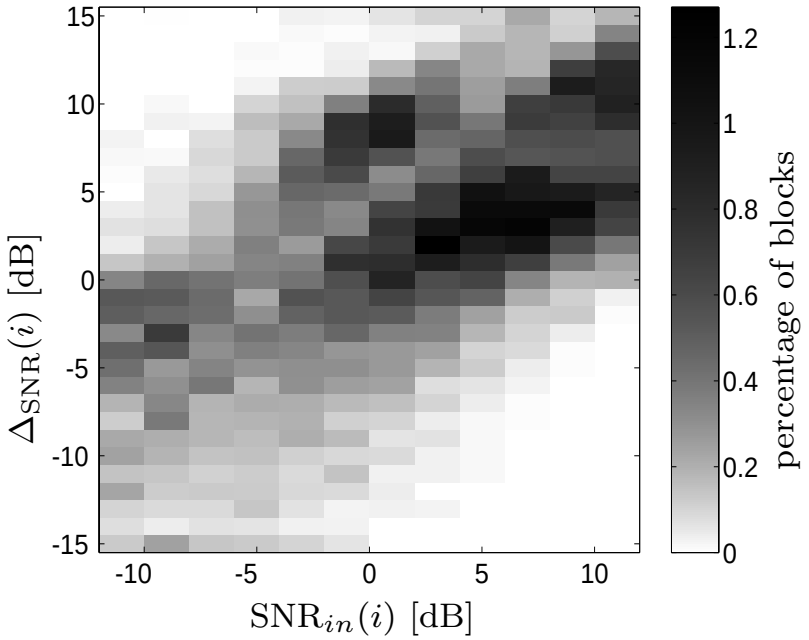

Figure 9: 2D-histogram over percentage of blocks with certain SNR$_{in}(i)$ achieving a certain $\Delta_{\mathrm{SNR}}(i)$ during speech periods.

In addition to displaying results of the proposed algorithm, results of conventional spectral subtraction algorithms are also presented applying a Hanning or rectangular window to the input samples block, $\mathbf{x}_L$.

The averaged (between blocks) sequences presented in Figures 10 and 11 are absolute mean averages of the output from the IFFT, $\overline{|y_N(n)|}$ (see Figure 1) for each sample within the blocks. The averaging is performed over blocks on the absolute value of each sample in the 256 long data block from the IFFT, so that

$$\overline{|y_N(n)|} = \frac{1}{I}\sum_{i=1}^{I}|y_N(n,i)| \tag{22}$$

where I is the number of blocks and n is the sample number in a block, i.e. $n \in [0, 255]$. This means that the effects of different choices of gain function are clearly visible, i.e. non-causal filtering, shorter and longer impulse responses, minimum phase or linear phase. As seen in the figures, the part that will be overlapped with the next block in the overlap-add function has much reduced sample values for the proposed method as compared to the conventional method. This effect leads to reduced interference between blocks.

In conventional spectral subtraction the block discontinuity effect can be reduced by employing a Hanning window, and introducing a 50 percent overlap. The drawback of this solution is the extra delay of $L/2 = 80$ samples. The proposed method reduces the discontinuity effect without introducing the delay. The new method imposes a causal property on the gain function introducing only a short delay, at most, $\frac{M-1}{2}$ samples. When using minimum phase filtering the delay is only a few samples.

When the sound quality of the output signal is the most important factor, a linear phase filter is often used. A good compromise if the processing delay must be short is to choose the minimum phase filter, although the computational complexity is higher as compared with the linear phase filter.

3.4 Adaptive Exponential Averaging of $G_M(f, i)$

The adaptive averaging of the gain function should provide lower variance when the signal is stationary. The main advantage is thereby the reduction of residual noise, i.e. musical tones in the background. The gain function with and without exponential averaging is presented in Figures 12 and 13, respectively. As can be seen in the figures, the variability of the signal is lower during noise only periods and also for low energy speech periods, when adaptive averaging is employed. The lower variability of the gain function gives less audible tonal artifacts in the output signal.

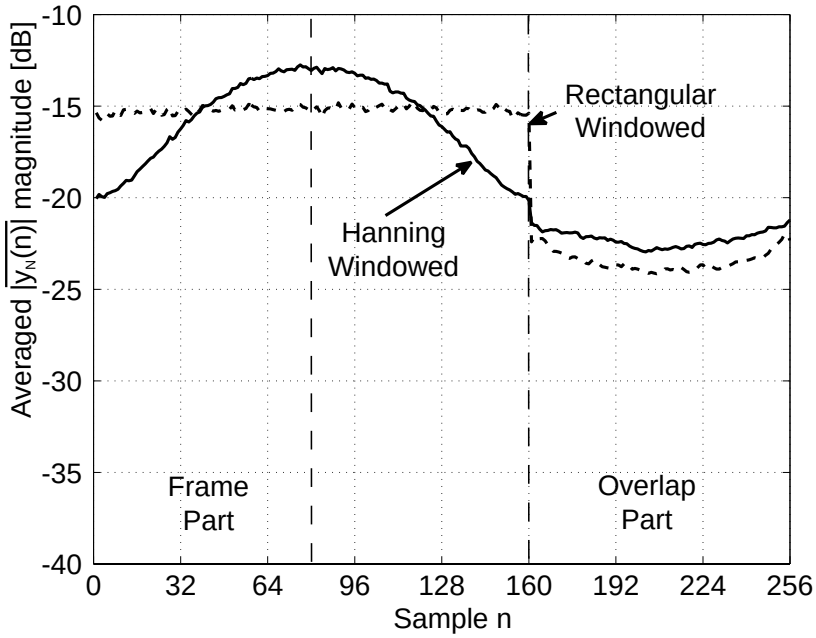

Figure 10: Mean absolute value output block, $\overline{|y_N(n)|}$, when using conventional spectral subtraction i.e. filtering with periodic convolution and a zero phase filter. The input time block has been windowed by a rectangular or Hanning window. The vertical lines show the borders for overlap between frames, at sample 80 when using Hanning window and at sample 160 when using rectangular window. The high energy samples in the block's overlap part will interfere with other block's frame part.

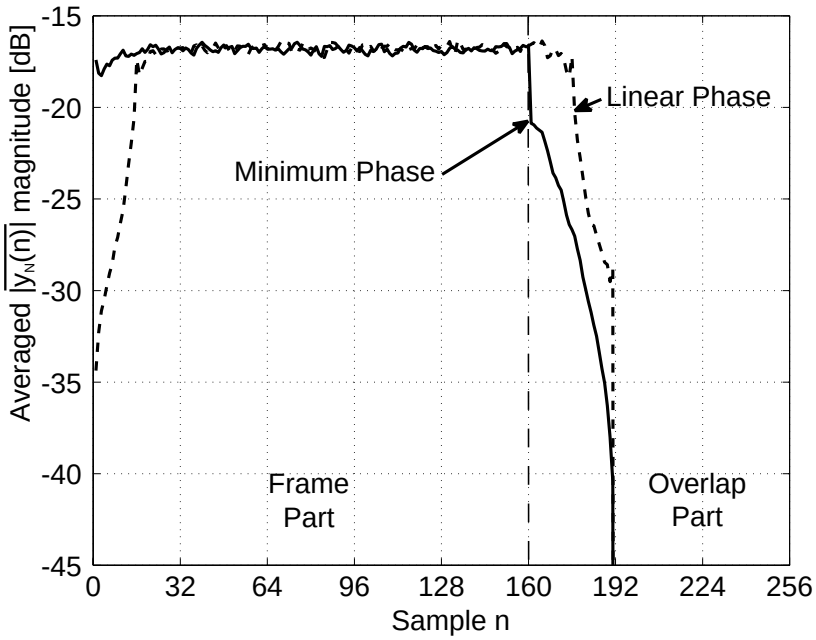

Figure 11: Mean absolute value output block, $\overline{|y_N(n)|}$, when using the proposed spectral subtraction, i.e. filtering with truly linear convolution and a linear phase or minimum phase filter. The vertical line shows the border for overlap between frames. The phase applied to the gain function makes this causal. The benefit with causal filters can be observed by the low level of the samples in the overlap part.

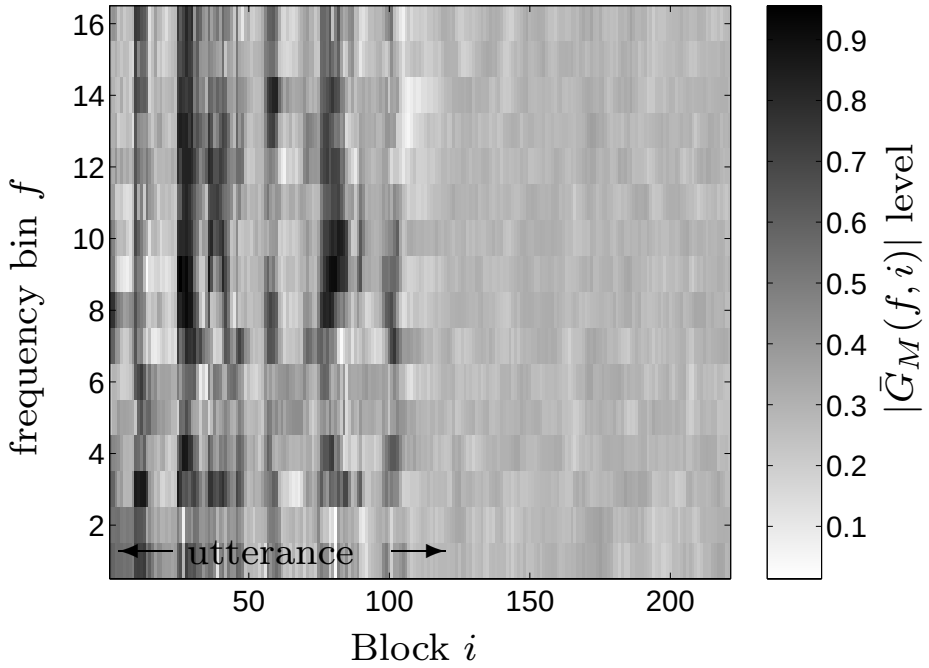

Figure 12: Absolute value of the gain function, $|\bar{G}_M(f,i)|$, with exponential averaging "on".

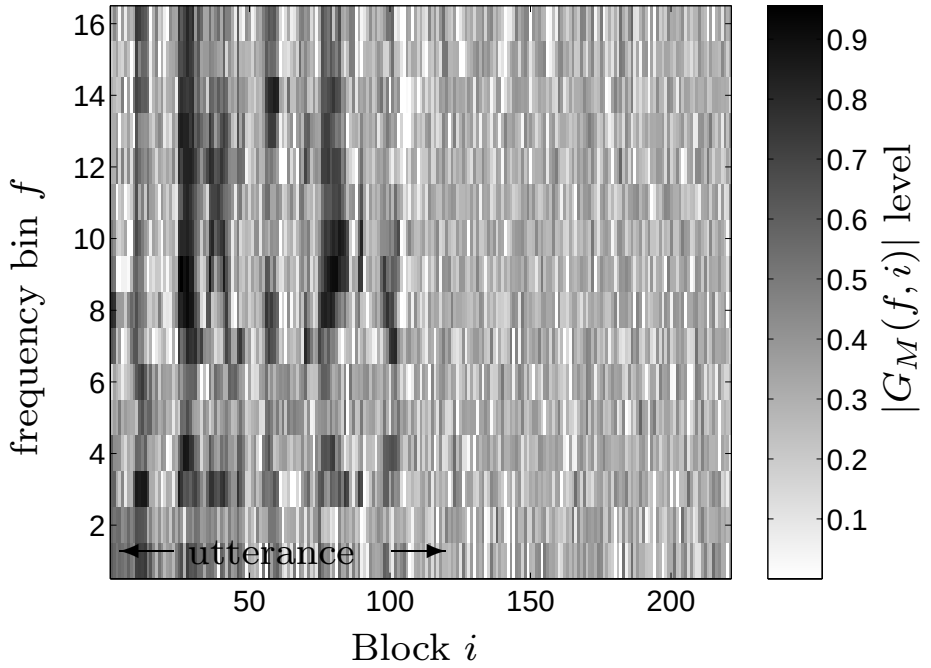

Figure 13: Absolute value of the gain function, $|G_M(f, i)|$, with exponential averaging "off".

3.5 Objective Sound Quality Measures

It is difficult to find relevant objective sound quality measures. There exist some objective sound distortion measures which may indicate sound quality, [11, 12]. The methods used here are the median of per-frame measures over all speech blocks. The per-frame measures are Itakura-Saito Distortion (IS), Log-Likelihood Ratio (LLR), Log-Area-Ratio (LAR) and Weighted Spectral Slope (WSS), all of which are described in Appendix A. For the spectral subtraction method presented we compare the quality of the input speech, $s(n)$, and the processed speech, $y_s(n)$, using these suggested distortion measures. When the processed signal is undistorted the quality measures are near zero, and the more distortion in the signal the higher the distortion measure, as expected. In Figure 14, the objective sound distortion measures are plotted, showing a peak in distortion for total input SNR:s in the range 0–10 dB. This result is due to the fact that the noise reduction is strongest in this range. When using conventional spectral subtraction methods it is not possible to obtain the processed speech separate from the noise since the total filtering of the input is not linear. Therefore, when comparing the quality to other spectral subtraction methods, the noise reduced speech, $y(n)$, is used for all the methods to get a fair comparison. See Table 1 for a comparison between the quality of the noisy speech input signal, $x(n)$, and the processed signals using conventional spectral subtraction, spectral subtraction using over-subtraction and noise floor [5], and the presented spectral subtraction. Most of the measures indicate that the presented method has an improved speech quality. The processed background noise distortion is low according to the objective measures, see Table 2. The objective quality measures indicate good sound quality and informal listening tests seem to confirm these results.

Method	IS	LLR	LAR	WSS
Input $x(n)$	0.22	0.18	2.85	39.5
Conventional SS	0.19	0.16	2.54	48.5
Over-sub with floor SS	1.53	0.47	4.19	67.0
Presented SS	0.32	0.15	2.44	41.5

Table 1: Objective sound quality measures for spectral subtraction methods during speech periods.

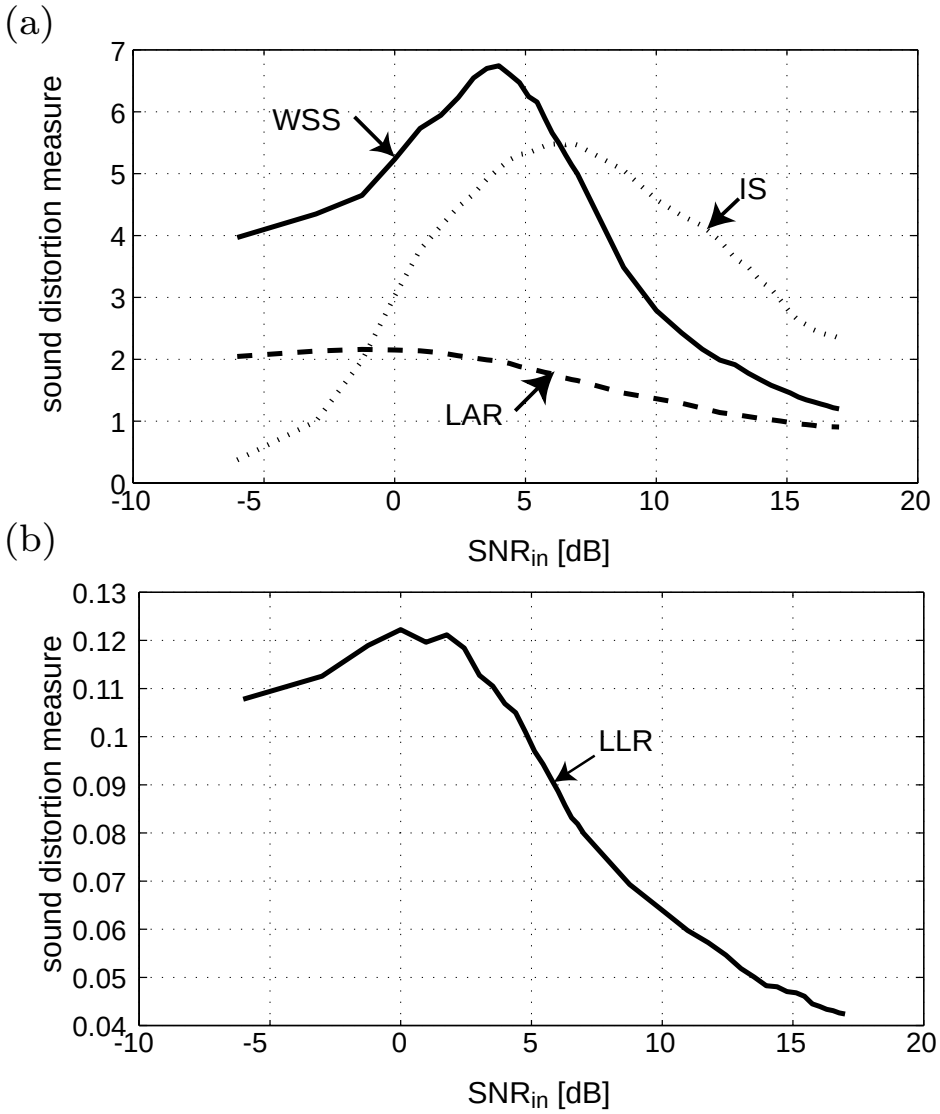

Figure 14: The objective sound distortion measures for the processed speech, $y_s(n)$, (a) WSS, IS and LAR measures, (b) LLR measure. The distortion measures are close to zero for clean speech and higher values indicate a greater distortion.

Method	IS	LLR	LAR	WSS
Conventional SS	75.4	0.23	2.90	34.0
Over-sub with floor SS	279	0.18	2.95	45.0
Presented SS	192	0.06	1.84	4.30

Table 2: Objective sound quality measures for spectral subtraction methods during noise only periods.

4 Conclusions & Further Work

An improved spectral subtraction method has been suggested and presented. This method provides a noise reduction procedure which functions well with arbitrary frame lengths, gives low residual noise, high quality speech, and low background noise artifacts, and introduces only a short delay. These are important properties when the noise reduction methods are integrated together with other speech enhancement methods and speech coders in real-time communication systems.

The method reduces the variability of the gain function – in this case a complex function – in two ways. First, the variance of the current block's spectrum estimate is reduced using the Bartlett method by trading frequency resolution for variance reduction. Second, an adaptive averaging of the gain function is used which is dependent on the discrepancy between the estimated noise spectrum and the current input signal spectrum estimate. The low variability of the gain function during stationary input signals gives an output with less tonal residual background noise, thus low noise distortion. The lower resolution of the gain function is also utilized to perform a truly linear convolution leading to improved sound quality. The sound quality is further enhanced by adding causal properties to the gain function.

The results show that the quality improvement can also be observed in the output block. The output blocks interfere less with other blocks when they are fitted with the overlap-add method giving improved sound quality. The low interference is due to the causal properties of the gain function which give low sample values in the overlap part of the frame. The output noise reduction is approximately 10 dB using the parameter choices used in this paper, making this method suitable for real-time noise reduction systems, e.g. in a mobile phone.

An extention that can be made to this work is to exploit a spectrum estimation method more similar to the human hearing, e.g. a Bark or Mel scale scheme.

5 Acknowledgment

We would like to thank Jim Rasmusson, formerly at Ericsson Mobile Communications, for his support of this research. We would also like to thank Timothy Samuels for proof-reading the manuscript. This work was supported by The Foundation for Knowledge and Competence Development and by Ericsson Mobile Communications AB, Sweden.

A Appendix: Objective Sound Quality Measures

The Itakura-Saito (IS) Distortion Measure is derived from the linear prediction (LP) coefficient vector, $\mathbf{a}_s(i)$, of the original clean speech frame and the processed speech coefficient vector, $\mathbf{a}_y(i)$, resulting in,

$$\mathrm{IS}(i) = \frac{\sigma_s^2(i)}{\sigma_y^2(i)} \frac{\mathbf{a}_y(i)\mathbf{R}_s(i)\mathbf{a}_y^T(i)}{\mathbf{a}_s(i)\mathbf{R}_s(i)\mathbf{a}_s^T(i)} + \log\left(\frac{\sigma_y^2(i)}{\sigma_s^2(i)}\right) - 1 \tag{23}$$

where $\sigma_s^2(i)$ and $\sigma_y^2(i)$ are the all-pole gains for the processed and clean speech, respectively, and $\mathbf{R}_s(i)$ denotes the input clean speech signal correlation matrix. Compared to the LLR measure this measure has the advantage of giving a zero response when estimating the distortion of two signals having equal LP coefficients and gains.

Log-Likelihood-Ratio (LLR) Measure is given by

$$\mathrm{LLR}(i) = \log\left(\frac{\mathbf{a}_y(i)\mathbf{R}_s(i)\mathbf{a}_y^T(i)}{\mathbf{a}_s(i)\mathbf{R}_s(i)\mathbf{a}_s^T(i)}\right). \tag{24}$$

This measure can be interpreted as the difference of the residuals when filtering the clean speech signal with the inverse LP coefficients filter of the clean speech as well as the processed speech.

Log-Area-Ratio (LAR) Measure is calculated by

$$\mathrm{LAR}(i) = \left|\frac{1}{P}\sum_{p=1}^{P}\left(\log\frac{1+r_{s,p}(i)}{1-r_{s,p}(i)} - \log\frac{1+r_{y,p}(i)}{1-r_{y,p}(i)}\right)^2\right|^{\frac{1}{2}} \tag{25}$$

where $r_{s,p}(i)$ and $r_{y,p}(i)$ are the p:th reflection coefficients of the input clean speech signal and of the processed signal, respectively.

The Weighted Spectral Slope (WSS) Measure is based on an auditory model in which a critical-bands filterbank is used to estimate the short-time

speech spectrum. The spectral slopes in each band are calculated. The distances between the spectral slopes of clean speech and the spectral slopes of processed speech are weighted according to each bands distance from a spectral peak or vally. The weighting function in the distortion measure put emphasize on formant placement. For a more precise definition, see [12]. An advantage of this measure is the aural model which gives a spectral distance incorporating perceptually meaningful frequency weighting.

References

[1] S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech and Sig. Proc.*, 27:113-120, 1979.

[2] M. Winberg, I. Claesson: "Spectral Subtraction with Extended Methods", *Research Report HK-R*, Aug. 1996.

[3] N. Virage: "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", *IEEE Trans. Speech and Audio proc.*, vol 7, no 2, pp 126-137, 1999.

[4] D. Tsoukalas, M. Paraskevas, J. Mourjopoulos: "Speech enhancement using psychoacoustic criteria", *Proc. IEEE ICASSP*, 359-362 vol.2, 1993.

[5] M. Berouti, R. Schwartz, J. Makhoul: "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE ICASSP*, 208-211, 1979.

[6] P. Händel: "Low-distortion spectral subtraction for speech enhancement", *Proc. Eurospeech 95*, vol. 2, pp 1549-1552, Spain, Sep., 1995.

[7] P.P. Vaidyanathan: "Multirate Systems and Filter Banks" *Prentice-Hall*, 1993.

[8] J. G. Proakis, D. G. Manolakis: "Digital Signal Processing; Principles, Algorithms, and Applications", *Macmillan*, Second Ed., 1992.

[9] A. V. Oppenheim, R. W. Schafer: "Discrete - time signal processing.", *Prentice-Hall*, Inter. Ed., 1989.

[10] European digital cellular telecommunications system (Phase 2), "Voice Activity Detection (VAD) (GSM 06.32)", *European Telecommunications Standards Institute*, 1994.

[11] J. H. L. Hansen, Bryan L. Pellom: "An effective quality evaluation protocol for speech enhancement algorithms", *Proc. ICSLP 98*, Sydney, Australia, Dec. 1998.

[12] S. R. Quackenbush, T. P. Barnwell, M. A. Clements: "Objective Measures of Speech Quality", *Prentice Hall*, Advanced Reference Series, 1988.

Part B

Dual-Microphone Spectral Subtraction

H. Gustafsson, S. Nordholm, I. Claesson, U. Lindgren, "Dual-Microphone Spectral Subtraction", Research Report 2000:2, University of Karlskrona/Ronneby, ISSN 1103–1581, February, 2000.

Dual-Microphone Spectral Subtraction

Harald Gustafsson, Sven Nordholm, Ingvar Claesson and Ulf Lindgren

Abstract

Mobile phones are constantly decreasing in size, thereby complicating the acoustical functionality. Signal processing methods can be used to partially mitigate this problem. In this paper we suggest a method which uses multiple spectral subtraction functions and two microphones, introducing only a short signal delay. The idea is to use spectral subtraction methods to extract the noise as well as the speech during a single time-frame. The environment background noise may not be stationary, thereby limiting the method to only employ short estimates of the background noise signal. Results are presented for experiments in various environments, showing a reduced noise level in the processed signal compared with the un-processed signal, and with preserved speech quality.

1 Introduction

A vital property of mobile phones is the speech quality experienced both by the far-end and the near-end user. There are several sources that obstruct a comfortable conversation, i.e. background noise, artificial delays of the speech, and missing parts of the speech and echoes. Since mobile phones are decreasing in size the acoustic functionality is central in retaining speech quality. When the distance from the microphone to the mouth is increased, the microphone picks up more of the background noise, which is disturbing to the far-end speaker. One way of reducing this disturbance, by improving the acoustic pick-up with a cavity, is to use a flip. The main drawback with flips is, however, that they are mechanically fragile and rather expensive to manufacture. A preferred solution is therefore often to use signal processing methods to improve the microphone signal. The present report presents a method to reduce the background noise introducing only a very short delay.

Mobile telephones are used in almost all environments. Therefore, the background noise can exhibit highly varying characteristics over time. A common method used to reduce background noise is spectral subtraction. Spectral subtraction employs two spectrum estimates: one estimate of the speech signal disturbed by a background noise signal spectrum, and an estimate of the background noise signal spectrum. These are combined to form an SNR-based (Signal to Noise Ratio) gain function [1, 2, 3] in order to reduce the background noise. The background noise amplitude spectrum is normally estimated during speech pauses. This implies that the background must contain a similar amplitude spectrum during the speech periods. Unfortunately, this assumption is not always true for many common background noise situations. To alleviate the problem of estimating the background noise a second microphone can be introduced, which picks-up more of the background noise.

Many ideas for dual-microphone speech enhancement depend upon delay differences of the two-microphone signals and thereby assume specific spatial conditions for the sound sources. A classic approach to enhance the speech with multiple microphones is beamforming [5], and evaluations of improved beamformers for speech enhancement can be found in [6, 7]. Beamforming methods use the phase information of the two microphone signals to selectively include a lobe in the desired direction and thereby increase the SNR of the speech signal. The performance of two microphone beamforming methods decreases when exposed to more than two sources [8]. Another approach with beamforming techniques is to explore the possibility of nulling the speech signal content of one of the microphone signals with the beamformer output, thereby obtaining a noise spectrum estimate which subsequently is used in a spectral subtraction method [8]. Adaptive Noise Cancellation (ANC) is yet

another classic approach for reducing the noise content in a signal [9]. ANC methods assume that the second microphone signal contains a noise signal without any speech content. The noisy-speech microphone signal is fed to an adaptive linear filter with the noise signal as the desired signal.

Other approaches to estimate the noise in spectral subtraction is to utilize the power spectrum and the cross spectrum estimates between the two signals [10]. This approach assumes low coherence in the two microphone signals for the background noise signals and a high coherence for the speech signals. Signal separation has also been suggested for enhancing the speech signal [11, 12]. This signal separation deconvolutes (point) sources which are mixed, by minimizing the square cross-correlation between the two microphone signals. However, other signal separation criteria also exist. The number of microphone signals limits the number of sources that can be separated, so that when two signals are used a maximum of two sources can be separated.

In this report it is suggested that the second microphone should be mounted more distant from the mouth, as in Figure 1, picking up the present frame background noise. This signal can be used to reduce the background noise in the primary microphone signal, thus facilitating the handling of short time stationary background noise. The second microphone signal will, however, also contain speech residues. In order to use this second microphone signal as a reliable noise estimate the speech signal must be suppressed. To do so the primary microphone signal is used since it contains a similar speech signal amplitude spectrum, with a better SNR. The SNR of the primary microphone signal can also be further improved by another spectral subtraction. This noise-reduced primary signal is used to form a spectral subtraction on the second microphone signal, thereby extracting the background noise signal. These two pre-processing spectral subtraction blocks are combined resulting in a background noise signal estimate.

A final spectral subtraction is used directly on the primary signal by employing the enhanced background noise amplitude spectrum estimate from the pre-processing.

In order to handle the energy level differences between the two microphone signals several adaptive subtraction factors are suggested. The subtraction factors control the amount of reduction of the unwanted signals. These factors give a trade-off between speech distortion and noise reduction. The suggested method exploits a frame-wise mean amplitude measure levels of the two microphone signals to estimate suitable subtraction factors.

A major challenge in spectral subtraction is to obtain low variance amplitude spectrum estimates. Such a single-microphone spectral subtraction algorithm is used in each of the three enhancement functions [2, 3]. This

method reduces the variability of the gain function by using Bartlett's spectrum estimation method [4] in order to reduce the variance of the amplitude spectrum estimates when calculating the gain function filter. The short-time stationarity assumption is limited to the duration of the present signal frame, in order to comply with the demands of the spectrum estimation method. The Bartlett spectrum estimation method has lower frequency resolution as compared to a conventional periodogram [4]. The calculated gain function and the input signal spectrum should be represented in an equal number of frequency bins to facilitate the filtering. The filter and signal are interpolated to a suitable length, exceeding the combined original lengths. A phase is also imposed on the filter. By doing so, a causal, purely linear convolution is performed. The variability of the gain function is further reduced by adaptive averaging. The adaptation is controlled by a discrepancy measure between the background noise and the noisy speech spectrum estimate [3].

2 Acoustical and Mechanical Setup

The two microphones can be mounted quite differently, for example on a mobile telephone or an earpiece. When selecting proper placement of the microphones, consideration has to be taken to the sound propagation, both from the speaker and the background noise sources. The locations of the microphones are important, and should comply with the requirement of having a stronger speech signal in one of the two microphone signals. The two microphone signals should also have similar amplitude spectral shapes. An obvious solution is to locate the primary microphone at the bottom of the device closer to the mouth, while the second microphone is placed at the top of the device, closer to the ear. The microphones should also be placed so that the risk is low that the user, by mistake, covers or shadows the microphones.

When the device is a small mobile telephone the secondary microphone is approximately located at the doubled distance from the mouth as compared to the primary microphone, see Figure 1. Since sound energy decreases with the distance from the source, the speech energy level ratios between the two microphone signals only varies slowly, although the individual energy levels may vary more quickly. This slow variation is motivated by the assumption that the microphone positions are fixed and the phone is held at approximately the same angle during a time frame. The energy level difference of the speech signal in the two microphone signals is in the interval of 7–11 dB. The energy levels of the background noise in the two microphone signals are approximately equal. The differences that occur are mainly due to microphone directionality.

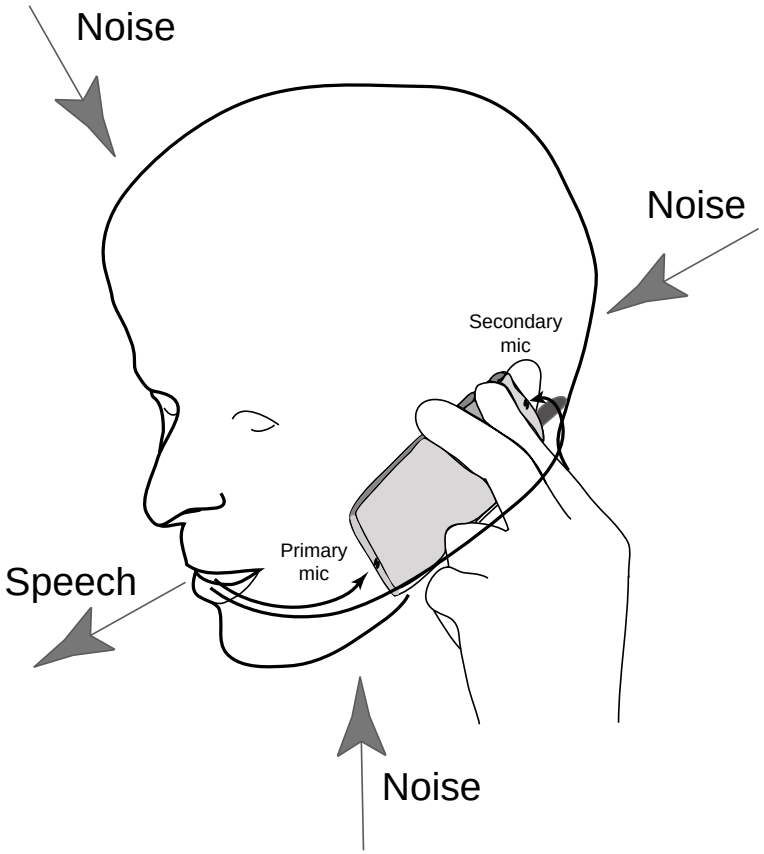

Figure 1: The two microphones are mounted at the top and bottom of a mobile phone. Typical distance from mouth to primary microphone is 7 cm, and to the secondary microphone 15 cm.

Typical spectra of the signals picked up by the microphones are shown in Figures 2 – 4. It can be seen in Figure 2 that the background noise spectra are similar in the two microphone signals for this outdoor recording even if the background mainly consists of multiple unwanted speakers, so-called babble noise. For the car recordings, the background noise has a small energy level difference, which increases for higher frequencies, as shown in Figure 3. For clean speech, the energy level difference is approximately 11 dB between the two microphone signals, see Figure 4.

When the background noise sources are closer to the telephone user there may be energy level differences of the received background noise signal in the two microphones.

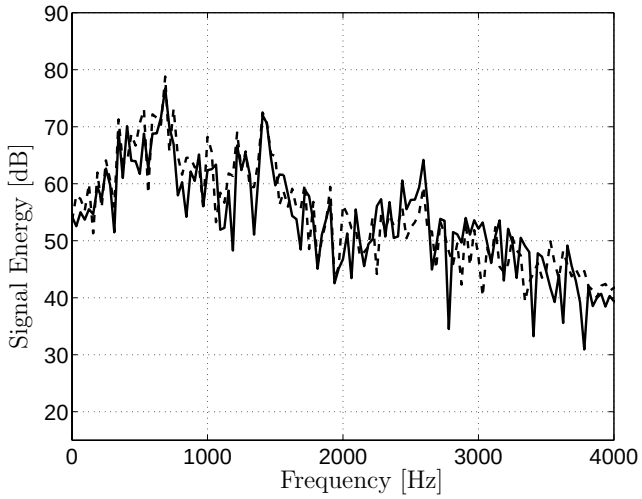

Figure 2: Babble noise power spectra recorded adjacent to an open-air cafeteria. Solid line: second microphone signal; dashed line: first microphone signal.

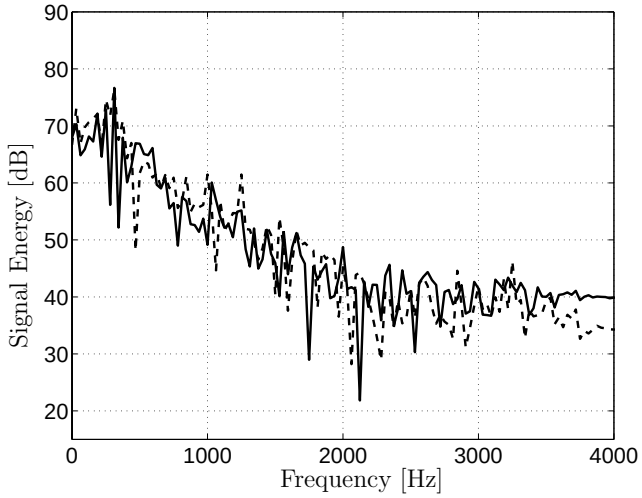

Figure 3: Noise power spectra recorded in a car. Solid line: second microphone signal; dashed line: first microphone signal.

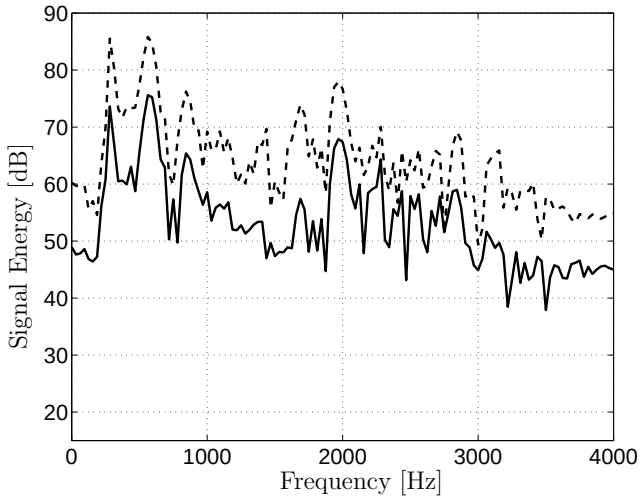

Figure 4: Speech power spectra recorded in a quiet anechoic chamber. Solid line: second microphone signal; dashed line: first microphone signal.

3 Dual-Microphone Noise Reduction Algorithm

The suggested method consist of three spectral subtraction blocks. The task of each block is to enhance speech or noise. First two spectral subtraction blocks are combined in a sequential manner to estimate the background noise amplitude spectrum, this is the pre-processing block. A third, alias final, spectral subtraction block uses the noise amplitude spectrum estimate, in order to enhance the primary microphone speech signal.

The primary microphone signal is denoted by $x_1(n)$ and the secondary microphone signal by $x_2(n)$. The signals contain both speech and background noise. The background noise signal is considered to be additive and uncorrelated with the speech signal. The speech signals are denoted by $s_1(n)$ and $s_2(n)$ in the primary and the secondary microphone signals, respectively. The background noise signals are likewise denoted by $n_1(n)$ and $n_2(n)$. The inputs are thus given by

$$x_1(n) = s_1(n) + n_1(n), \tag{1}$$

$$x_2(n) = s_2(n) + n_2(n). \tag{2}$$

The presented method works frame-wise in the frequency-domain. To enable a frequency interpretation of an FFT-transformation from the time-domain to the frequency-domain the signals are assumed to be short-time stationary. The stationarity should last as long as the duration of a block of samples, i.e. L samples. The blocks of samples are defined as the vectors

$$\mathbf{x}_{1,L}(i) = \begin{bmatrix} x_1(Li) & x_1(Li+1) & \dots & x_1(L(i+1)-1) \end{bmatrix}, \tag{3}$$

and

$$\mathbf{x}_{2,L}(i) = \begin{bmatrix} x_2(Li) & x_2(Li+1) & \dots & x_2(L(i+1)-1) \end{bmatrix}. \tag{4}$$

The blocks, $\mathbf{x}_{1,L}(i)$ and $\mathbf{x}_{2,L}(i)$, are divided into subblocks of length M

$$\begin{aligned} &\mathbf{x}_{1,L}(i;m) = \\ &\begin{bmatrix} x_1(Li+Mm) & x_1(Li+Mm+1) & \dots & x_1(Li+M(m+1)-1) \end{bmatrix}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} &\mathbf{x}_{2,L}(i;m) = \\ &\begin{bmatrix} x_2(Li+Mm) & x_2(Li+Mm+1) & \dots & x_2(Li+M(m+1)-1) \end{bmatrix}. \end{aligned} \tag{6}$$

The short-time amplitude spectrum estimate of the primary microphone signal is estimated by using the Bartlett method [4]

$$\hat{P}_{x_1,M}(f,i) = \sqrt{\frac{M}{L}\sum_{m=0}^{\frac{L}{M}-1} |\mathcal{F}_M\{\mathbf{x}_{1,L}(i;m), f\}|^2}, \tag{7}$$

where $f \in [0, M-1]$ is a discrete variable enumerating the M frequency bins, and $\mathcal{F}$ is the FFT operation. The secondary microphone signal amplitude spectrum estimate $\hat{P}_{x_2,M}(f,i)$ is analogously defined from the secondary microphone signal block $\mathbf{x}_{2,L}[n](i)$. The Bartlett method calculates the periodograms on the $\frac{L}{M}$ sub-blocks and averages the ensemble of periodograms. This process results in an amplitude spectrum estimate which has a lower variance as compared to the full-block length (L) periodogram. The resulting amplitude spectrum has also a reduced frequency resolution. However, the reduced variance is here a trade-off with respect to frequency resolution.

3.1 Mean Amplitude Measure

The short term amplitude spectrum, $\hat{P}_{x_1,M}(f,i)$, is defined in Equation (7). In order to control the three involved spectral subtraction algorithms a frame-wise Mean Amplitude Measure, MAM, is defined as

$$A_{1,x}(i) = \frac{1}{M}\sum_{f=0}^{M-1} \hat{P}_{x_1,M}(f,i). \tag{8}$$

This MAM is used to control the different subtraction factors k_r, k_n, and k_s.

3.2 Spectral Subtraction

The Spectral Subtraction (SS) method is presented in detail in Figure 5 and Figure 6. The main ideas are similar as for a single microphone noise reduction, see Appendix A. The gain functions, $G_{(\cdot),M}(f,i)$, for the three blocks are formed by both the estimated amplitude spectrum of the input signal, $\hat{P}_{x_{(\cdot)},M}(f,i)$ see Equations (1) – (7), and the estimated amplitude spectrum of the contaminating signal, $\hat{P}_{y_{(\cdot)},M}(f,i)$, and are given by

$$G_{r,M}(f,i) = \left(1 - k_r \cdot \frac{|\hat{P}_{y_n,M}(f,i-1)|^a}{|\hat{P}_{x_1,M}(f,i)|^a}\right)^{\frac{1}{a}}, \tag{9}$$

$$G_{n,M}(f,i) = \left(1 - k_n \cdot \frac{|\hat{P}_{y_r,M}(f,i)|^a}{|\hat{P}_{x_2,M}(f,i)|^a}\right)^{\frac{1}{a}}, \tag{10}$$

and

$$G_{s,M}(f,i) = \left(1 - k_s \cdot \frac{|\hat{P}_{y_n,M}(f,i)|^a}{|\hat{P}_{x_1,M}(f,i)|^a}\right)^{\frac{1}{a}}, \tag{11}$$

where a is the spectrum exponent, $k_{(\cdot)}$ is a subtraction factor controlling the amount of suppression, the subscript r denotes the *rough* speech gain function, n denotes the *noise* gain function, and s denotes the final *speech* gain function. The Bartlett method in Equation (7) is used to obtain an amplitude spectrum estimate with lower variance. Having a low variance in the spectrum estimate results in a low variability of the gain function. This, in turn, results in less artifacts in the filtered output signal. When the input signal is long-time stationary the gain function $G_{s,M}(f,i)$ can be averaged adaptively [3]. This is done to further reduce the residual artifacts, mainly during non-speech periods. The averaging adaptation is dependent on the spectral discrepancy between the primary input amplitude spectrum estimate, $\hat{P}_{x_1,M}(f,i)$, and the extracted noise amplitude spectrum estimate $\hat{P}_{y_n,M}(f,i)$. The smaller the spectral discrepancy, the longer the averaging time employed.

The gain functions correspond to non-causal time-varying filters if zero-phase is assumed. To obtain a causal filter, a minimum phase [13], or a

linear phase [4], characteristics is imposed on the gain function, resulting in $\tilde{G}_{(\cdot),M}(f,i)$.

Since the spectral subtraction technique is frame-based and uses FFTs, frame effects must be considered. An FFT corresponds to a critically sampled filter bank [14]. A circular convolution (which comes from the FFT and IFFT operations) results in discontinuities between frames but can be avoided by using correct lengths of the filter length M and data frame length L. To obtain purely linear convolution the corresponding time sequence of the gain function, $\tilde{G}_{(\cdot),M}(f,i)$, and the spectrum estimate, $\hat{X}_{(\cdot),L}(f,i)$, must yield a length which is shorter than the FFT length, i.e. $L+M<N$ [13]. An interpolated gain function, $\tilde{G}_{(\cdot),M\uparrow N}(f,i)$, is used with the same number of FFT bins as the interpolated (zero-padded) input spectrum, $\hat{X}_{(\cdot),L\uparrow N}(f,i)$. Although the filter and spectrum have N frequency bins they are of order of M and L, respectively. An enhanced signal without periodicity artifacts can be obtained as

$$Y_{r,N}(f,i)=\tilde{G}_{r,M\uparrow N}(f,i)\cdot X_{1,L\uparrow N}(f,i), \tag{12}$$

$$Y_{n,N}(f,i)=\tilde{G}_{n,M\uparrow N}(f,i)\cdot X_{2,L\uparrow N}(f,i), \tag{13}$$

and

$$Y_{s,N}(f,i)=\tilde{G}_{s,M\uparrow N}(f,i)\cdot X_{1,L\uparrow N}(f,i). \tag{14}$$

where the subscript r denotes *rough* speech estimate, n denotes *noise* estimate, and s denotes processed *speech* estimate.

The last spectrum estimate, $Y_{s,N}(f,i)$, is transformed to the time-domain by using an IFFT resulting in the sample block

$$\mathbf{y}_{s,N}(i)=\mathcal{F}_N^{-1}\{Y_{s,N}(f,i),f\}. \tag{15}$$

The resulting time-domain signal is achieved using the overlap-add method. Since the output from the pre-processing blocks, namely the spectra $Y_{r,N}(f,i)$ and $Y_{n,N}(f,i)$, are of length N, a decimation in frequency resolution is performed, giving the amplitude spectra $P_{y_r,M}(f,i)$ and $P_{y_n,M}(f,i)$. The decimation is necessary for the calculation of the short-length gain functions. The decimation is made by first applying an N length IFFT on the output spectra $Y_{r,N}(f,i)$ and $Y_{n,N}(f,i)$. This is followed by employing the Bartlett method of sub-block length M yielding

$$\mathbf{y}_{r,N}(i)=\mathcal{F}_N^{-1}\{Y_{r,N}(f,i),f\}, \tag{16}$$

$$\hat{P}_{y_r,M}(f,i) = \sqrt{\frac{M}{N}\sum_{m=0}^{\frac{N}{M}-1} |\mathcal{F}_M\{\mathbf{y}_{r,N}(i;m), f\}|^2}, \tag{17}$$

$$\mathbf{y}_{n,N}(i) = \mathcal{F}_N^{-1}\{Y_{n,N}(f,i), f\}, \tag{18}$$

$$\hat{P}_{y_n,M}(f,i) = \sqrt{\frac{M}{N}\sum_{m=0}^{\frac{N}{M}-1} |\mathcal{F}_M\{\mathbf{y}_{n,N}(i;m), f\}|^2}. \tag{19}$$

3.3 Total Algorithm Structure

The total algorithm is as follows: the microphone closest to the mouth contains signal with the better Signal-to-Noise Ratio (SNR). This signal should be processed resulting in an even further enhanced speech signal, i.e. better SNR although with some speech distortion. When handling noise within a single frame a good estimate of that frame's noise amplitude spectrum is plausible. This is achieved by using the second microphone signal, although this signal contains a weaker version of the speech signal as well. By using a combination of three spectral subtraction schemes as in Figure 6, this difficulty can be overcome.

The speech pre-processing spectral subtraction function, SS_r, in Figure 6 uses a frame, $\mathbf{x}_{1,L}(i)$, of the primary signal and an amplitude spectrum estimate of the noise enhanced secondary signal computed in the previous frame, $P_{y_n,M}(f,i-1)$, to form a rough spectrum estimate of the speech signal, $Y_{r,N}(f,i)$. This is a rough estimate since it is calculated with the noise spectrum estimate of the previous frame and the subtraction factor is set at a high level. By setting the subtraction factor high, a strong noise reduction is achieved, which also distorts the speech signal. Still, some of the noise remains and artifacts are introduced. The noise pre-processing spectral subtraction, SS_n, uses the rough extracted speech amplitude spectrum estimate $P_{y_r,M}(f,i)$ and a block, $\mathbf{x}_{2,L}(i)$, of the secondary signal to form a spectrum estimate of the background noise signal for the current frame, $Y_{n,N}(f,i)$. The secondary microphone signal is used since the speech signal energy level is lower in this signal, which simplifies suppression of the speech signal part. When the two spectral subtraction blocks are tuned, SS_r will give a rough high-SNR speech estimate and SS_n will give a high-NSR noise estimate. If

one of the blocks' performance deteriorates, the other will follow, due to the coupling.

The final speech enhancement procedure, SS_s, uses the current background noise amplitude spectrum estimate, $P_{y_n,M}(f,i)$, and again block, $\mathbf{x}_{1,L}(i)$, of the primary signal to obtain a noise reduced speech spectrum, $Y_{s,N}(f,i)$. The background noise amplitude spectrum estimate is block-wise updated for each new block reflecting changes of the background, even during speech periods.

3.4 Processing Delay

Communication systems have a specified maximum delay that has to be fulfilled. The delay must be kept as low as possible in speech communication to prevent unnatural pauses and stuttering. When the signal frame length is matched to the mobile phone voice coder frame length, the proposed method can work on the same frame of samples as the voice coder, which means that *no* extra delay is introduced. The introduced delay is the computation time of the noise reduction plus the delay in the gain function filter in SS_s. The computation delay is included in the total delay since the frame of samples must be collected that much in advance to be able to send it at a predetermined time. When the minimum phase is imposed on the gain function the filtering delay is less than half a millisecond.

4 Adaptive Control Mechanism

The noise suppression method suggested in this report controls the suppression by the three subtraction factors k_r, k_n and k_s. The levels of these parameters must change according to the sound environment for the mobile telephone. The factors should regulate the level of suppression of the contaminating signal and also compensate for the different amplitude levels of both the background and the speech signal spectra in the two microphone signals.

The frame-wise mean amplitude measure, MAM, in the microphone signals are denoted by $A_{1,x}(i)$ and $A_{2,x}(i)$ for the primary and the secondary microphone signal, respectively. The frame-wise MAM of the speech signal in the primary and secondary microphone signals are denoted by $A_{1,s}(i)$ and $A_{2,s}(i)$, respectively, and the corresponding background noise MAM are denoted by $A_{1,n}(i)$ and $A_{2,n}(i)$, respectively.

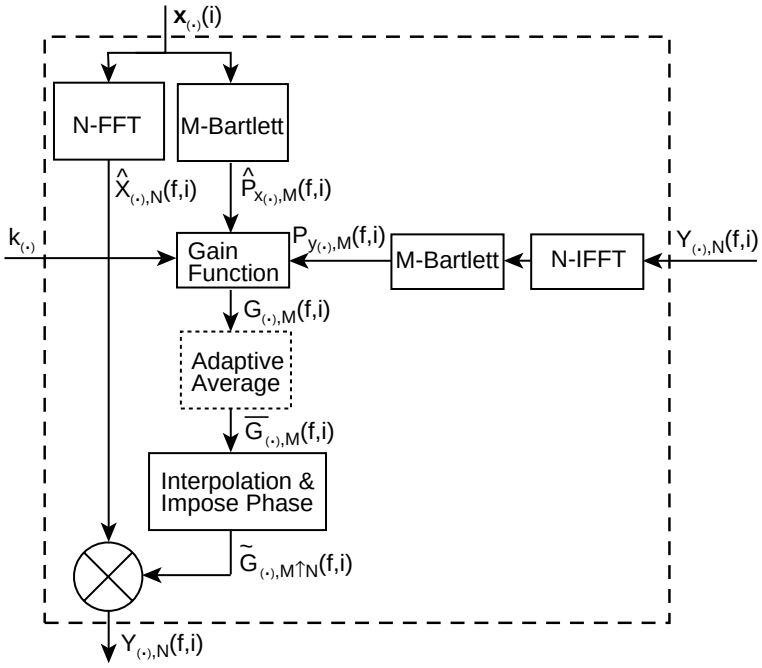

Figure 5: The spectral subtraction function; the adaptive averaging is optional and only used in SS_s.

4.1 Interpretation of the Subtraction Factors

The speech pre-processing spectral subtraction block should have a strong noise reduction and thereby a higher level of the subtraction factor, k_r, is used. The subtraction factor should be set to the level at which the SS_r block, see Figure 6, results in a speech signal with lowest noise level. A low level of residual artifacts is not the primary goal here. The parameter choice of k_r must also take into consideration amplitude differences of the background signal in the two microphone signals. When the background amplitude in the secondary microphone signal is higher than the level in the primary microphone, k_r should decrease, hence

$$k_r \propto \frac{A_{1,n}(i)}{A_{2,n}(i)}, \tag{20}$$

where $\propto$ denotes a dependence between the operands, so that when one operand increases or decreases the other follows.

The noise spectral subtraction block in Figure 6 , SS_n, is used to extract the noise part in the secondary microphone signal. The subtraction factor k_n controls how much of the speech signal that should be suppressed. Since

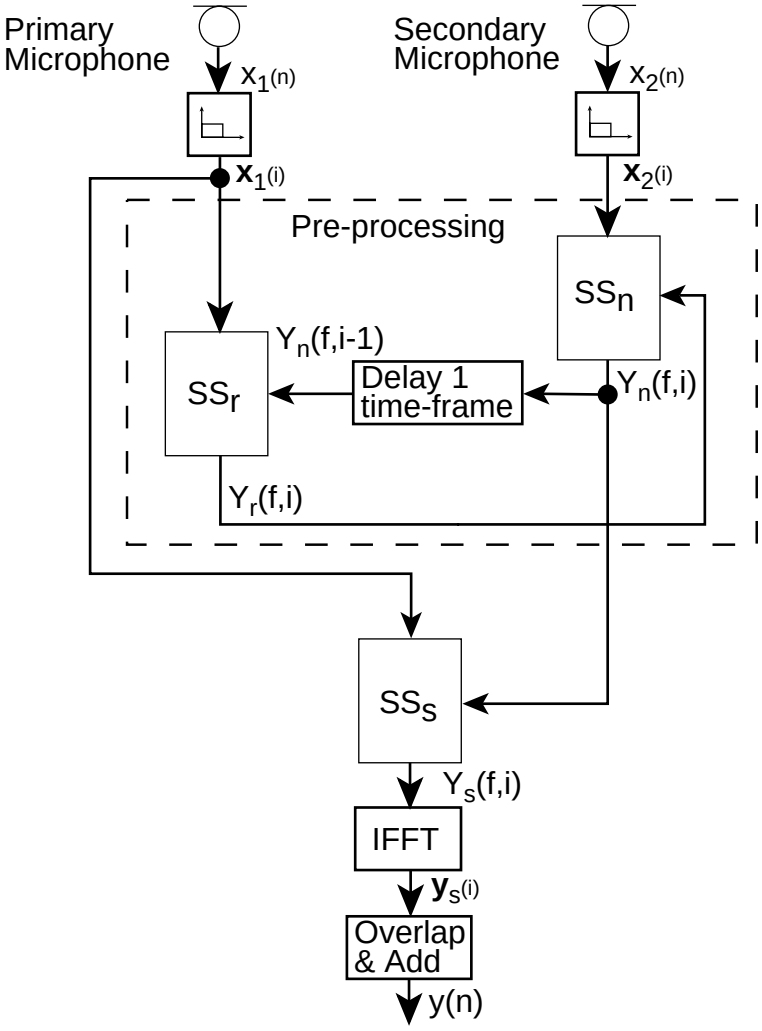

Figure 6: The noise reduction procedure consists mainly of three spectral subtraction functions, which are executed in the order SS_r, SS_n, SS_s.

the speech signal in the primary microphone signal has a higher energy level than in the secondary microphone signal, k_n should be selected accordingly, hence

$$k_n \propto \frac{A_{2,s}(i)}{A_{1,s}(i)}. \tag{21}$$

The resulting noise estimate should only contain a small residual of the speech signal, preferably no speech signal at all, since remains of the desired speech signal will be detrimental to the ultimate speech enhancement procedure, SS_s, thus lowering the quality of the output.

The final spectral subtraction block, SS_s, is controlled much the same way as SS_r. The difference is that the noise suppression is deemphasized in favor of the speech distortion. This generally implies that the subtraction factor, k_s, is lower than k_r.

4.2 Adaptive Control Mechanism

A central part of the process is taking into consideration the MAM of the speech and noise when selecting subtraction factors. Inspired by the spectral subtraction equation a method is suggested such that the amplitudes in the two microphones are equalized.

The subtraction factor is derived by using a method inspired from Equation (9)

$$\hat{A}_{1,s}(i) \approx \left(1 - k_r(i) \cdot \frac{\hat{A}_{2,n}(i-1)}{A_{1,x}(i)}\right) \cdot A_{1,x}(i) \tag{22}$$

where the extracted MAMs are distinguished from the real data measured MAMs by a hat above the parameter. In Equation (22), the exponent parameter a has been selected to one, and the spectra have been replaced by the MAMs, $\hat{A}_{1,s}(i)$ and $\hat{A}_{2,n}(i-1)$, which are the MAMs of the output from the speech, SS_s, and the noise, SS_n, pre-processor, respectively. Solving Equation (22) for the direct subtraction factor, $k_r(i)$, gives

$$k_r(i) \approx \frac{A_{1,x}(i) - \hat{A}_{1,s}(i-1)}{\hat{A}_{2,n}(i-1)}. \tag{23}$$

Equation (23) make use of the MAM of the previous frame and the present frame, which can result in a mismatch of amplitude levels. To reduce the mismatch between frames in the calculation Equation (23) is reformulated, $\hat{A}_{1,s}(i-1)$ is approximated by $A_{1,x}(i)(1 - \bar{g}_{r,M}(i-1))$, this yields

$$\tilde{k}_r(i) = \frac{A_{1,x}(i)(1 - \bar{g}_{r,M}(i-1))}{A_{2,x}(i)\bar{g}_{n,M}(i-1)} \cdot \kappa_r \tag{24}$$

where κ_r is also introduced, a fixed multiplication factor setting the overall noise reduction level, and

$$\bar{g}_{r,M}(i) = \frac{1}{M} \sum_{f=0}^{M-1} G_{r,M}(f,i), \tag{25}$$

$$\bar{g}_{n,M}(i) = \frac{1}{M} \sum_{f=0}^{M-1} G_{n,M}(f,i). \tag{26}$$

The gain functions are limited to $0 \leq G_{r,M}(f,i) \leq 1$ and $0 \leq G_{n,M}(f,i) \leq 1$. The summation of gains over frequency gives an overall estimation of speech to noise ratio and noise to speech ratio of the frame, respectively. Equation (24) depends on the ratio of the noise levels in the two microphone signals. Besides κ_r Equation (24) only adjust for differences in amplitude between the two microphones. The subtraction factor $\tilde{k}_r(i)$, increases during speech periods. This is a suitable behavior since a stronger noise reduction is desired during these periods. If a subtraction factor with similar level during speech and non-speech periods is used, the obtained SNR improvement is too low during speech periods. Much of the remaining noise during speech periods is not perceived since it is masked by the hearing, but when the extracted speech is used to suppress the speech in the secondary microphone signal the remaining noise will deteriorate the noise estimate.

To reduce the variability of $\tilde{k}_r$ to a reasonable range a limited and averaged subtraction factor is introduced

$$\bar{k}_r(i) = \frac{1}{J_r + 1} \sum_{j=0}^{J_r} \begin{cases} k_{r,max}(i), & \tilde{k}_r(i-j) > k_{r,max}(i) \\ \tilde{k}_r(i-j), & k_{r,min} < \tilde{k}_r(i-j) < k_{r,max}(i) \\ k_{r,min}, & \tilde{k}_r(i-j) < k_{r,min} \end{cases} \tag{27}$$

where $J_r + 1$ is the number of averaged subtraction factors, $k_{r,min}$ is the minimum allowed $\bar{k}_r$, and $k_{r,max}(i)$ is the maximum allowed $\bar{k}_r$ calculated by

$$k_{r,max}(i) = \min([\bar{k}_r(i), \bar{k}_r(i-1) \ldots, \bar{k}_r(i-\Delta_r)]) + r_r \tag{28}$$

where the maximum is set by an offset, r_r, to the minimum $\bar{k}_r(i)$ found during the last Δ_r frames. The maximum $k_{r,max}(i)$ is used to prevent too high a subtraction level during speech periods, and to decrease the fluctuations of the gain function. Parameter Δ_r should be large enough to cover the last noise only period at least partially. The averaged subtraction factor is subsequently used in the spectral subtraction, see Equation (9), instead of the direct subtraction factor k_r.

The parameter $\bar{k}_s(f,i)$ is derived in the same way as $\bar{k}_r(i)$ except that it is calculated for each frequency bin separately,

$$\tilde{k}_s(f,i) = \frac{|\hat{P}_{x_1,M}(f,i)| \cdot (1 - G_{r,M}(f,i))}{|\hat{P}_{x_2,M}(f,i)| \cdot G_{n,M}(f,i)} \cdot \kappa_s, \tag{29}$$

$$\bar{k}_s(f,i) = \frac{1}{J_s+1}\sum_{j=0}^{J_s}\begin{cases} k_{s,max}(i), & \tilde{k}_s(f,i-j) > k_{s,max}(i) \\ \tilde{k}_s(f,i-j), & k_{s,min} < \tilde{k}_s(f,i-j) < k_{s,max}(i) \\ k_{s,min}, & \tilde{k}_s(f,i-j) < k_{s,min} \end{cases}, \tag{30}$$

$$k_{s,max}(i) = \min([\bar{k}_s(f,i), \bar{k}_s(f,i-1)\ldots,\bar{k}_s(f,i-\Delta_s)]) + r_s, \quad f \in [0,1,\ldots,M-1] \tag{31}$$

where $\bar{k}_s(f,i)$ is the subtraction factor at discrete frequencies $f \in [0,1,\ldots,M-1]$. The frequency dependent subtraction factor is motivated by the fact that the transfer function between the two microphone signals is also frequency dependent. This frequency dependence is varying over time due to movement of the mobile phone. A frequency dependence could also be used for the two first subtraction factors, but in order to reduce computational complexity we have refrained from doing so since speech quality is most important in only the final spectral subtraction function.

Even though the subtraction factor is calculated in each frequency band it is smoothed over frequencies to reduce its variability giving

$$\bar{\bar{k}}_s(f,i) = \frac{1}{V}\sum_{v=-\frac{V-1}{2}}^{\frac{V-1}{2}} \bar{k}_s([f+v]_0^M, i) \tag{32}$$

where V is the odd length of a rectangular smoothing window, and $[\cdot]_0^M$ is an interval restriction of the frequency at 0 and M. The subtraction factor $\bar{\bar{k}}_s(f,i)$ smoothed in both frequency and frame directions is used in Equation (11) instead of the direct subtraction factor.

The noise pre-processor subtraction factor is different since it controls the amount of speech signal that should be removed from the second microphone signal. This subtraction factor is derived from

$$\hat{A}_{2,n}(i) \approx \left(1 - k_n(i)\cdot\frac{\hat{A}_{1,s}(i)}{A_{2,x}(i)}\right)\cdot A_{2,x}(i). \tag{33}$$

this expression is inspired from Equations (10) and (13). In Equation (33) the spectra have been replaced by the frame-wise MAM. Solving Equation (33) for the direct subtraction factor, $k_n(i)$, gives

$$k_n(i) \approx \frac{A_{2,x}(i) - \hat{A}_{2,n}(i-1)}{\hat{A}_{1,s}(i)}\cdot\kappa_n, \tag{34}$$

where an overall speech reduction level, κ_n, is also introduced. Without explicitly using the amplitude measures of the pre-processed signals a more robust control is accordingly obtained by

$$\tilde{k}_n(i) = \frac{A_{2,x}(i)(1 - \bar{g}_{n,M}(i-1))}{A_{1,x}(i)\bar{g}_{r,M}(i)} \cdot \kappa_n. \tag{35}$$

Subtraction factor $\tilde{k}_n(i)$ depends on the ratio between the speech levels in the two microphone signals.

To reduce the variability and to limit $\tilde{k}_2$ to an allowed range the exponentially averaged subtraction factor is obtained,

$$\bar{k}_n(i) = \beta_n \cdot \bar{k}_n + (1 - \beta_n) \cdot \begin{cases} k_{n,max}, & \tilde{k}_n(i) > k_{n,max} \\ \tilde{k}_n(i), & k_{n,min} < \tilde{k}_n(i) < k_{n,max} \\ k_{n,min}, & \tilde{k}_n(i) < k_{n,min} \end{cases} \tag{36}$$

where β_n is the exponential averaging constant, $k_{n,max}$ is the maximum allowed $\bar{k}_n$, and $k_{n,min}$ is the minimum allowed $\bar{k}_n$. The averaged subtraction factor is then used in the spectral subtraction Equation (10) instead of the direct subtraction factor k_n.

There are numerous possibilities for controlling the subtraction factors. The different methods can be used separately or in conjunction to compensate for weaknesses found in other methods.

5 Evaluation

The evaluation is made using both quantitative and qualitative measures. The decrease of noise level achieved is presented by means of noise reduction and SNR improvement measures. Quality is not that straightforward to evaluate. We have chosen to employ four degradation measures and performed informal listening tests.

The suggested dual-microphone algorithm is compared with a one microphone algorithm [3] which is outlined in Appendix A. The latter algorithm has a primary microphone signal input only.

The evaluations are performed on input signals where different background noise signals are added to clean speech signals. The selected noise environments are a car coupé, near an open-air cafeteria, and next to a city street with traffic. The speech signals were recorded in a quiet anechoic chamber. The same recording setup was used for all the recordings. The microphones were fixed at the bottom and top of a solid piece of plastic, a small mobile phone dummy. The dummy was held in a typical position for

	min	max
k_r	0.5	—
k_n	0.1	0.6
k_s	0.5	—

Table 1: The limits for the subtraction factors.

κ_r	1.0
κ_n	0.8
κ_s	0.9

Table 2: Multiplication factors for calculation of the subtraction factors.

all recordings. A multi-channel measuring DAT recorder was used to gather signals from the microphones. The sound signals were anti-aliased, down-sampled to 8 kHz, and subsequently filtered by a telephone bandwidth filter, i.e. 300–3400 Hz.

5.1 Parameter Choices

The frame length was $L = 160$ to comply with the GSM voice coder. A suitable filter length is chosen to $M = 32$. This gives an FFT length of minimum $N = 256$ since a radix-2 method is used. An amplitude spectral subtraction is chosen by setting $a = 1$, since this produces a stronger noise reduction with less perceived speech degradation.

The adaptive control mechanism of the subtraction factors has many parameters. Most of the parameters are used to transform and limit the factors to a required range. The limitations are only invoked to prevent abnormal behavior, see Table 1 for the levels selected. The frame maximum level for subtraction factors $\tilde{k}_r(i)$ and $\tilde{k}_s(i)$ are dependent on added quantities, r_r and r_s, respectively, which are added to the found minima during the Δ_r and Δ_s most recent frames. The parameters are set to $r_r = 0.5$, $r_s = 0.3$, $\Delta_r = 100$, and $\Delta_s = 100$ which corresponds to two seconds at 8 kHz. The multiplication factors selected, κ_r, κ_n, and κ_s, are given in Table 2.

The averaging of the subtraction factors $\bar{k}_r(i)$ and $\bar{k}_s(i)$ is done over $J_r = 3$ and $J_s = 3$ recent frames, respectively. The exponential averaging of $\bar{k}_n(i)$ is controlled by the exponential averaging constant $\beta_n = 0.6$. Finally, the smoothing of $\bar{\bar{k}}_s(f, i)$ over frequency is set by the rectangular window length $V = 5$.

5.2 Noise Reduction

Once all parameters are selected and calculated based on data for each frame. The noise reduction algorithm can be applied to perform an effective evaluation of the actual noise reduction and SNR improvement achieved on the inputs separately. Since the gain function is a filter, although time-varying between frames, the primary speech signal and background noise signal can be filtered separately for evaluation purposes. The pre-calculated gain function calculated on the combined input data is used to process the individual signals. The sum of the outputs will result in the same output signal as if the noisy speech were filtered,

$$\begin{aligned} y_s(n) = h(x_1(n), n) = h(s_1(n) + n_1(n), n) = \\ h(s_1(n), n) + h(n_1(n), n) = y_{s,s}(n) + y_{s,n}(n) \end{aligned} \tag{37}$$

where $y_{s,s}(n)$ and $y_{s,n}(n)$ are the processed speech signal and background noise signal, respectively, and $h(\bullet, n)$ denotes a linear time-varying operator. It is now feasible to calculate the block energy for the input speech signal, $p_{1,s}(i)$, input background noise signal, $p_{1,n}(i)$, processed speech signal, $p_{s,s}(i)$, and processed background noise signal, $p_{s,n}(i)$. To evaluate the performance, block-wise SNR and noise reduction measures are defined as

$$\Delta_{SNR}(i) = \frac{p_{s,s}(i)}{p_{s,n}(i)} \cdot \frac{p_{1,n}(i)}{p_{1,s}(i)} = \frac{SNR_{out}(i)}{SNR_{in}(i)} = SNR_{out}(i)[\text{dB}] - SNR_{in}(i)[\text{dB}] \tag{38}$$

which is the Signal-to-Noise Ratio improvement for block i. The SNR before the spectral subtraction is denoted by $SNR_{in}(i)$ and the SNR after processing is denoted by $SNR_{out}(i)$,

$$SNR_{in}(i) = \frac{p_{1,s}(i)}{p_{1,n}(i)}, \tag{39}$$

$$SNR_{out}(i) = \frac{p_{s,s}(i)}{p_{s,n}(i)}. \tag{40}$$

The SNR measures are only valid during speech frames. Finally, the Noise Ratio, $NR(i)$, is defined as

$$NR(i) = \frac{p_{1,n}(i)}{p_{s,n}(i)}. \tag{41}$$

The achieved noise ratio during noise-only periods is higher as compared to speech periods. During noise-only periods, all the frequency bands can

be muted to a low level, but during speech periods the speech signal must be able to pass without being severely disturbed. This also has the effect that more noise passes through the filter during speech periods. Human hearing has a masking effect which makes low level sounds, close in both time and frequency to a high level sound, inaudible. The spectral subtraction algorithm reduces the noise in the low energy speech frequency bands but lets noise and speech pass in the higher energy speech frequency bands. This is perceived as general lower noise level when combined with the lower noise level during speech pauses. If noise reduction had been applied only during speech pauses it would have been perceived as a speech degradation.

The combined noise and speech evaluation signals are used as input signals to the dual-microphone noise reduction algorithm. The noise reduction during speech periods is measured and a histogram of the percentage of frames achieving different noise reductions is presented in Figure 7. As can be seen, the noise reduction is approximately 10 dB during speech pauses. The SNR improvement during speech periods is dependent upon the input SNR of the primary microphone signal. Figure 8 displays the percentage of frames with a certain SNR improvement dependent on the input SNR. The measured SNR improvement is less then 2 dB. Still no background noise level difference can be heard in the processed signal comparing speech and speech-pause periods. The speech signal masks some of the noise signal, so that an overall lower noise level is achieved.

5.3 Speech Quality Evaluation

The purpose of noise reduction is to maintain the speech quality of the input signal with a reduced background noise level. The speech quality can be evaluated both by objective and subjective means. Subjective measures are used, since objective methods only partly indicate the perceived speech quality.

5.3.1 Objective Degradation Measures

The objective speech degradation measures used in this report are presented in [15], [16] and Appendix B. The Log-Area-Ratio (LAR), Log-Likelihood-Ratio (LLR), and Itakura-Saito (IS) measures make use of the Linear Prediction Coefficients (LPC) to calculate the distortion. LPC analysis is often used for modelling speech production. The Weighted-Spectral-Slope (WSS) measure incorporates human perception in the calculations. All the methods estimate the displacement of the formants in their respective models. The degradations are tabulated in Table 3. As can be observed the input and

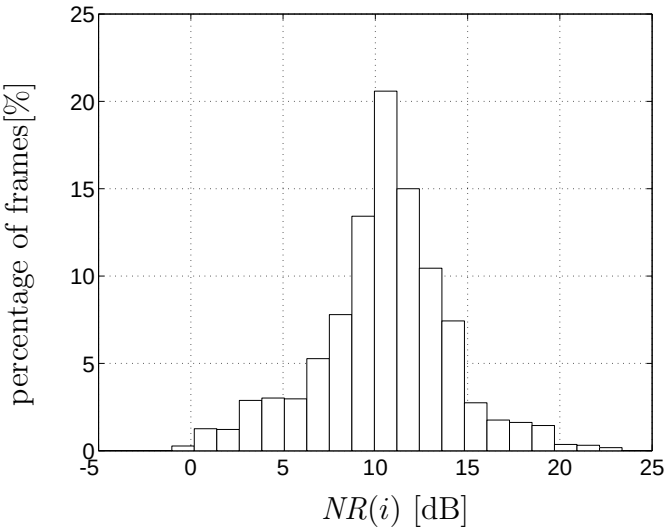

Figure 7: Histogram of the percentage of frames achieving a certain noise reduction during speech pauses when the dual-microphone noise reduction method is applied.

the output signals have approximately the same values, including the single microphone approach [3]. The single microphone approach has only one input, the primary microphone signal. The two rows at the bottom of Table 3 shows the degradation measures of the speech signal when it is filtered alone by the pre-calculated gain function. These values are lower than when the noise also is included, showing that it is mostly the remaining noise that contributes to the objective speech degradation measures.

5.3.2 Informal Listening Tests

Even though the objective measure does not indicate a difference between the dual-microphone approach and the single microphone approach, informal listening tests show that a difference can be heard. The dual-microphone approach is more uniform in quality. The single microphone approach has larger quality differences. When comparing the processed signals with the input signals the processed signals are considered to be better since the speech only exhibits a small degradation and the noise level is notably lower.

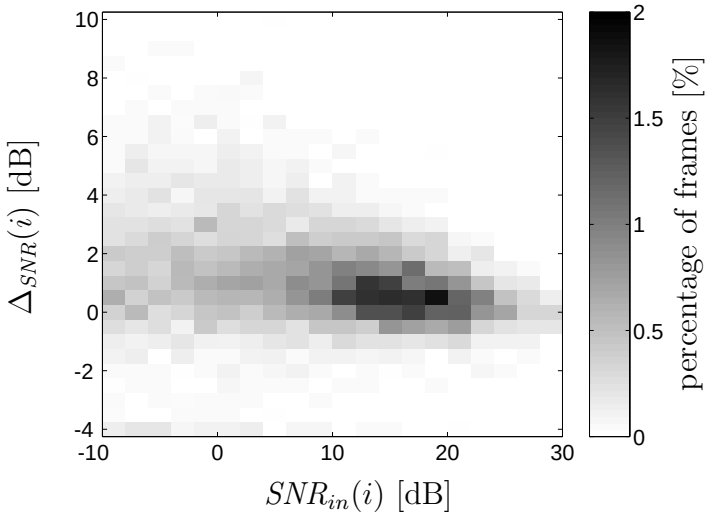

Figure 8: 2D-histogram of the percentage of frames achieving a certain SNR improvement at a certain input SNR during speech periods when the dual-microphone noise reduction method is applied. Even though the measured SNR improvement is low, no noise level difference between noise and speech periods can be heard. The small but essential noise reduction performed during speech periods gives the impression of a continuous, low noise level.

5.4 Filtering Delay

The noise reduction filtering operation delays the signal. When a linear phase is imposed on the filter, the delay will be fixed to $\frac{M-1}{2} = 15.5$ samples. When instead a minimum phase filtering is used the delay will be different for each frequency band. The delay can be characterized by means of the group delay. The group delay measures the delay of the envelope of a narrow band signal. Therefore, the influence that the filter, $G_{s,M\uparrow N}(f,i)$, has on the output signal, $y_s(n)$, is presented in Figure 9, as a histogram of the group delay for all frequency bands. Only frames containing speech signal are evaluated since the low-energy noise frames do not affect the perceived delay in the system. The observed delay is in the range of 0–4 samples corresponding to a delay of less than 0.5 ms. The mean delay is 0.4 samples. A causal minimum phase filter can most certainly have a negative group delay in the stop-bands which can be observed by noting that a small percentage of the frequency bands shows a negative delay.

		LAR	LLR	IS	WSS
Including the speech degradation effect of the noise	Input $x_1(n)$	3.5	0.36	0.54	30
	Dual-mic $y_s(n)$	*3.5*	*0.36*	0.66	*31*
	Single mic $y(n)$	*3.5*	0.37	*0.58*	32
Excluding the speech degradation effect of the noise	Input $s_1(n)$	0	0	0	0
	Dual-mic $y_{s,s}(n)$	*1.2*	*0.031*	*0.59*	*0.98*
	Single mic $y_{\hat{s}}(n)$	1.3	0.047	0.65	1.5

Table 3: The results with the four different objective degradation measures. The signals in the table are evaluated against the clean primary speech signal, $s_1(n)$. Lower speech degradation values in the table indicate better speech quality. The lower half of the table is calculated for speech signals only, processed by the pre-calculated gain functions.

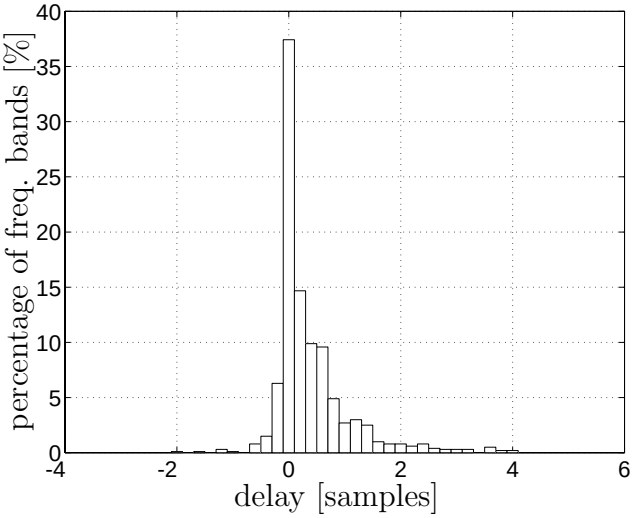

Figure 9: Histogram over all frequency bands and frames which yield a certain delay.

6 Conclusions

A dual-microphone noise reduction method has been proposed for use in mobile telephony. The method works well with short-time stationary input signals, gives low residual noise and high quality speech, and introduces only a short delay. These are important features in real-time handheld communication systems that may be used in complex sound environments.

The results show that it is possible to continually estimate the contaminating background signal's amplitude spectrum with reliability. The amplitude spectrum is used to calculate the noise suppression filter. When the filter is applied to the input signal, an output signal SNR improvement of 0–2 dB during speech periods is achieved together with a noise reduction of 10 dB during speech pauses. The delay of the processed signal is less than half a millisecond. These results indicate that the method is suitable for noise reduction in real-time systems, for example handheld mobile telephones.

7 Acknowledgments

This work was supported by The Foundation for Knowledge and Competence Development and by Ericsson Mobile Communications AB, Sweden. We thank Timothy Samuels for his careful proof reading of the report.

A Single Microphone Spectral Subtraction

The single microphone spectral subtraction algorithm is used as an integral part in the dual-microphone noise reduction method. The algorithm is outlined in Figure 10. Spectral subtraction relies upon the assumption that the background noise signal has an almost constant magnitude spectrum and the speech signal is short-time stationary. Furthermore, the background noise is considered additive and uncorrelated to the speech signal. Let $s(n)$, $w(n)$ and $x(n)$ represent the speech signal, noise signal and noisy speech signal, respectively, so that

$$x(n) = s(n) + w(n). \tag{42}$$

The short-time power spectral density relation is thus

$$R_x(f,i) = R_s(f,i) + R_w(f,i) \tag{43}$$

where $f \in [0, M-1]$ is a discrete variable enumerating the frequency bins and i is a time block index. The spectral subtraction method works in a

block-based fashion. The short-time spectral density is estimated by using a Bartlett method,

$$\hat{R}_{x,M}(f,i) = \sum_{n=0}^{\frac{L}{M}} |\mathcal{F}_M\{\mathbf{x}_L[n \cdot M, \ldots, (n+1) \cdot M - 1](i), f\}|^2 \qquad (44)$$

where $\mathbf{x}_L[n](i)$ is a vector containing the i:th block of L data samples, enumerated by n, and $\mathcal{F}$ is the FFT operation. The Bartlett method is used to get a spectrum with low variance and a reduced frequency resolution. For convenience, the magnitude spectrum estimate is defined as $\hat{P}_{x,M}(f,i) = \sqrt{\hat{R}_{x,M}(f,i)}$. The background noise magnitude spectrum can be estimated over a longer time frame by

$$\bar{P}_{w,M}(f,i) = \begin{cases} \mu\bar{P}_{w,M}(f,i-1) + (1-\mu)\hat{P}_{x,M}(f,i) & \text{, noise only} \\ \bar{P}_{w,M}(f,i-1) & \text{, speech and noise} \end{cases} \qquad (45)$$

where μ is the exponential averaging time constant. The speech pauses are detected by a Voice Activity Detector, VAD.

The low resolution spectrum estimates are used in the calculation of an SNR-based gain function,

$$G_M(f,i) = \left(1 - k \cdot \frac{\bar{P}^a_{w,M}(f,i)}{\hat{P}^a_{x,M}(f,i)}\right)^{\frac{1}{a}}, \qquad (46)$$

where a controls which power of magnitude spectral subtraction is to be used, k is the subtraction factor regulating the amount of noise reduction applied. In order to reconstruct a gain function that matches the number of FFT bins, N, the gain function is interpolated from the shorter gain function, $G_M(f,i)$, to form $G_{M\uparrow N}(f,i)$. Although $G_{M\uparrow N}(f,i)$ has N frequency bins, the corresponding impulse response is only of length M. We utilize the lower resolution of $G_{M\uparrow N}(f,i)$ to accomplish a truly linear filtering. Causality is imposed on the gain function by a linear or minimum phase. These novel properties are introduced on the gain function to facilitate improved speech quality as compared to other spectral subtraction methods. The resulting output is obtained by using overlap-add and an inverse FFT of

$$Y_N(f,i) = G_{M\uparrow N}(f,i)X_{L\uparrow N}(f,i). \qquad (47)$$

Another benefit with the novel method is the short delay introduced in the noise-reduced signal. When a minimum phase is imposed on the gain function the delay is only a few samples.

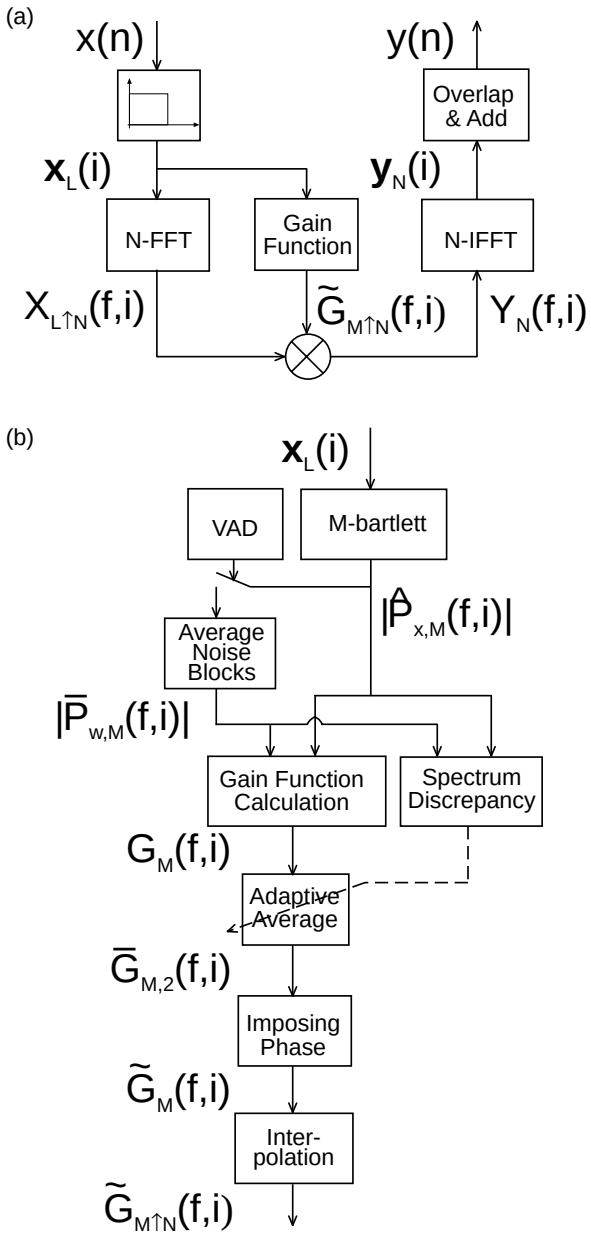

Figure 10: (a) Outline of the improved spectral subtraction algorithm. (b) A detailed view of the gain function calculation in (a), consisting of the parts facilitating the new causal truly linear filtering and adaptive exponential averaging.

B Objective Sound Quality Measures

The Itakura-Saito (IS) Distortion Measure is derived from the linear prediction (LP) coefficient vector, $\mathbf{a}_s(i)$, of the original clean speech frame and the processed speech coefficient vector, $\mathbf{a}_y(i)$, resulting in,

$$IS(i) = \frac{\sigma_s^2(i)}{\sigma_y^2(i)} \frac{\mathbf{a}_y(i)\mathbf{R}_s(i)\mathbf{a}_y^T(i)}{\mathbf{a}_s(i)\mathbf{R}_s(i)\mathbf{a}_s^T(i)} + \log\left(\frac{\sigma_y^2(i)}{\sigma_s^2(i)}\right) - 1 \tag{48}$$

where $\sigma_s^2(i)$ and $\sigma_y^2(i)$ are the all-pole gains for the processed and clean speech, respectively, and $\mathbf{R}_s(i)$ denotes the input clean-speech signal correlation matrix. Compared to the LLR measure, this measure has the advantage of giving a zero response when estimating the distortion of two signals having equal LP coefficients and gains.

The Log-Likelihood-Ratio (LLR) Measure is given by

$$LLR(i) = \log\left(\frac{\mathbf{a}_y(i)\mathbf{R}_s(i)\mathbf{a}_y^T(i)}{\mathbf{a}_s(i)\mathbf{R}_s(i)\mathbf{a}_s^T(i)}\right). \tag{49}$$

This measure can be interpreted as the difference of the residuals when filtering the clean speech signal with the inverse LP-coefficient-filter of the clean speech as well as the processed speech.

The Log-Area-Ratio (LAR) Measure is calculated by

$$LAR(i) = \left|\frac{1}{P}\sum_{p=1}^{P}\left(\log\frac{1+r_{s,p}(i)}{1-r_{s,p}(i)} - \log\frac{1+r_{y,p}(i)}{1-r_{y,p}(i)}\right)^2\right|^{\frac{1}{2}} \tag{50}$$

where $r_{s,p}(i)$ and $r_{y,p}(i)$ are the p:th reflection coefficients of the input clean speech signal and of the processed signal, respectively.

The Weighted-Spectral-Slope (WSS) Measure is based on an auditory model in which a critical-band filterbank is used to estimate the short-time speech spectrum. The spectral slopes in each band are calculated. The distances between the spectral slopes of clean speech and the spectral slopes of processed speech are weighted according to each band's distance from a spectral peak or valley. The weighting function in the distortion measure put emphasis on formant placement. See [16] for a more precise definition. An advantage of this measure is the aural model which gives a spectral distance incorporating perceptually meaningful frequency weighting.

References

[1] S. F. Boll: Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech and Sig. Proc.*, 27:113-120, 1979.

[2] H. Gustafsson, S. Nordholm, I. Claesson: "Spectral Subtraction Using Correct Convolution", *Proc. of ICSV6*, Copenhagen, Jul., 1999.

[3] H. Gustafsson, S. Nordholm, I. Claesson: Spectral subtraction, truly linear convolution and a spectrum dependent adaptive averaging method. *Submitted to IEEE Trans. Speech and Audio proc.*, Jun., 1999.

[4] J. G. Proakis, D. G. Manolakis: Digital Signal Processing; Principles, Algorithms, and Applications. *Macmillan*, Second Ed., 1992.

[5] L.J. Griffiths, C.W. Jim: "An alternative approach to linearly constrained adaptive beamforming", IEEE Trans. Antennas & Propag., Vol. AP-30, No. 1, pp. 27-34, Jan., 1982.

[6] S. Nordebo, S. Nordholm, B. Bengtsson, I. Claesson: "Noise Reduction Using an Adaptive Microphone Array in a Car- A Speech Recognition Evaluation", pp. 204, 16-18 in proc. 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, 1993.

[7] R.J.M. van Hoesel, G.M. Clark: "Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients", *Journal of the Acoustical Society of America*, Vol. 97, No. 4, p. 2498-503, Apri., 1995.

[8] K. Hack Yoon, F. Asano, Y. Suzuki, T. Sone: "Speech enhancement based on amplitude estimation with two-channel beamformer", *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol E79 A, No.12, p. 2151–8, Dec., 1996.

[9] B. Widrow, J.R Glover, J.M. Mcool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, R.C. Goodin: "Adaptive noise canceling: Principles and applications", Proc. IEEE, Vol. 63, pp. 1692-1716, 1975.

[10] M. Dorbecker, S. Ernst: "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation", *EUSIPCO*, Vol. 2, p. 995–8, Italy, 1996.

[11] H. Sahlin, H. Broman: "Separation of real-world signals", *EURASIP Signal Processing*, Vol. 64, No. 1, p. 103–13, Jan., 1998.

[12] F. Ehlers, H. G. Schuster: "Blind separation of convolution mixtures and an application in automatic speech recognition in a noisy environment", *IEEE Trans. on Signal Processing*, Vol. 45, No. 10, p. 2608–12, Oct., 1997.

[13] A. V. Oppenheim, R. W. Schafer: Discrete - time signal processing. *Prentice-Hall*, Inter. Ed., 1989.

[14] P.P. Vaidyanathan: "Multirate Systems and Filter Banks" *Prentice-Hall*, 1993.

[15] J. H. L. Hansen, Bryan L. Pellom: "An effective quality evaluation protocol for speech enhancement algorithms", *Proc. ICSLP 98*, Sydney, Australia, Dec. 1998.

[16] S. R. Quackenbush, T. P. Barnwell, M. A. Clements: "Objective Measures of Speech Quality", *Prentice Hall*, Advanced Reference Series, 1988.

Part C

Noise Suppression Employing Short Delay Processing

This part is published as: H. Gustafsson, S. Nordholm, I. Claesson, "Noise Suppression Employing Short Delay Processing", Proc. of ISPACS '99, Phuket, December 1999.

Noise Suppression Employing Short Delay Processing

Harald Gustafsson Sven Nordholm and Ingvar Claesson

Abstract

Many telecommunication systems allow only a short delay in the terminal. This restriction stems from the fact that a delay will disturb the conversation. Therefore only a short delay is permissible in noise reduction processing. The well-known spectral subtraction method has been modified to introduce only a short delay and a variance reduced gain function. This modification is accomplished by viewing the spectral subtraction as an FFT filter design method which designs a time-varying filter for each block of data. The design will, however, only give the amplitude function, thus in order to obtain a causal filter a phase must be imposed. The spectral subtraction will not provide any phase information hence a minimum phase is imposed on the filter. The designed filter is transformed to the time-domain and the actual noise reduction filtering will be performed in the time-domain with only a short delay. The proposed method reduces the noise by 10 dB with a maximum processing delay of 7 samples.

1 Introduction

The more mature the telecommunication market becomes the higher quality of service that will be requested. One quality parameter is the level of background noise and another is delay. The noise reduction capabilities have increased in block-based systems like the digital mobile phone systems. Office handsfree and conference telephones are sometimes used in disturbing background noise environments. In these situations it would be beneficial to apply noise reduction processing. One problem is that the allowed delay in PSTN-phone systems can be as short as 2 ms, that is 16 samples when using the normal telephone sampling rate of 8 kHz. One common method used for noise suppression is spectral subtraction [3]. The spectral subtraction is a block-based method which operates in the frequency-domain. This means that the calculated gain function is present after a buffer delay and a calculation delay. Some disadvantages using the conventional spectral subtraction are the so-called "musical tones" and block delay. The suggested method computes the time-variable filter in the frequency-domain, transforms it to the time-domain, and imposes a minimum phase, hence a short delay. The calculated gain function is a real and positive function, thus corresponding to a non-causal filter; the imposed minimum phase yields a causal filter. The highly varying gain function in conventional spectral subtraction, which results in musical tones, is avoided by the methods suggested in [1] and [2]. These suggested improvements are necessary to obtain a satisfactory result. The new idea in this paper has been to view the gain function calculation performed by the spectral subtraction and the multiplication of the input signal and the gain function as separate processes. A multiplication by the gain function in the frequency-domain corresponds to a convolution in the time-domain. Further, to avoid the block delay in spectral subtraction the filtering in the time-domain is performed on a sample-by-sample basis. The filter weights can be calculated using previous data. This means that if the processing of the input signal, $x(n)$, should be performed for the same data as used in the gain function calculation, an initial delay Δ must be inserted, see Figure 1. When the buffer length is set sufficiently short the delay is not required since the calculated filter will reflect the actual input signal, $x(n)$. This property can degrade the performance compared with a block-based method, but the benefits from a short delay will overshadow any disadvantages. Short delay makes it feasible to use noise reduction in more applications such as conference telephony and analogue hands-free telephony.

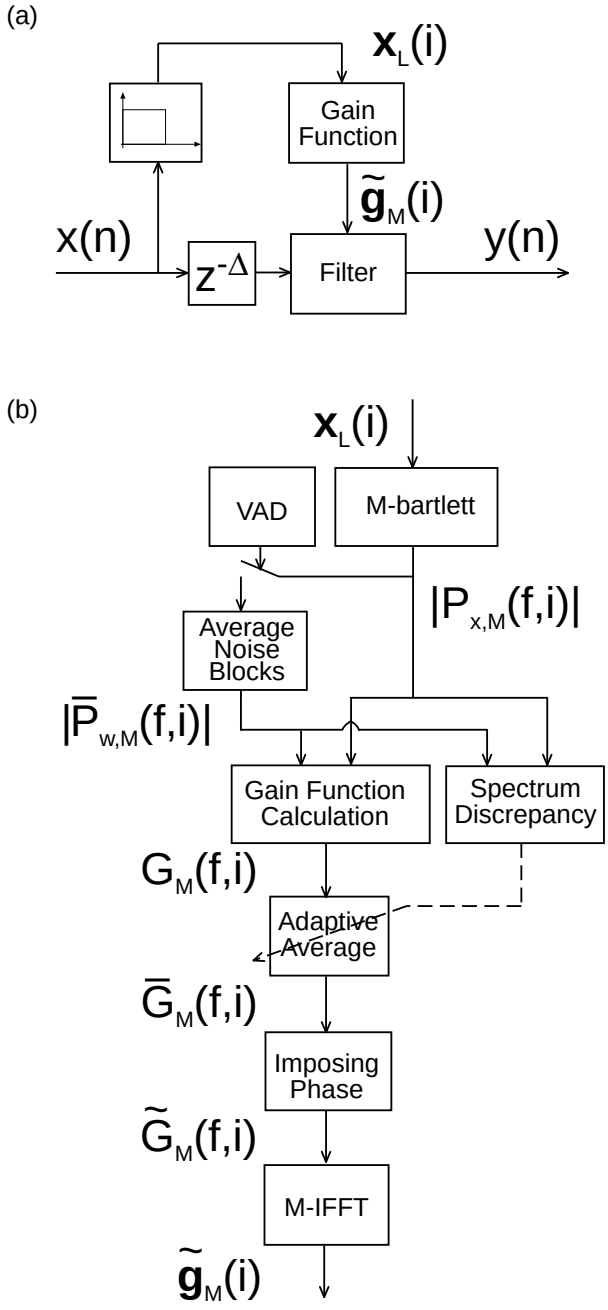

Figure 1: (a) Total structure of the new short delay noise reduction algorithm. (b) A detailed view of the gain function in (a), consisting of the parts facilitating the causal minimum phase filtering and adaptive exponential averaging.

2 Short delay noise reduction

Spectral subtraction assumes that the background noise signal has a "constant" magnitude spectrum and the speech signal is short time stationary, i.e. the magnitude spectrum varies slowly compared to the block length. Further, the noise signal is additive. Let $s(n)$, $w(n)$, and $x(n)$ represent the speech signal, the noise signal, and the noisy speech signal, respectively, such that

$$x(n) = s(n) + w(n). \tag{1}$$

The short-time power spectral density

$$R_x(f,i) = R_s(f,i) + R_w(f,i) \tag{2}$$

where $f \in [0, N-1]$ is a discrete variable corresponding to one frequency bin, and i is a block index. This short-time spectral density can be estimated by using Bartlett's method

$$\hat{R}_{x,M}(f,i) = \frac{M}{L} \sum_{p=0}^{\frac{L}{M}-1} \left| \mathcal{F}\{\mathbf{x}_{L,p}(i)\} \right|^2 \tag{3}$$

where $\mathbf{x}_{L,p}(i)$ is a non-overlapping partition p of length M of the i:th L-length input signal, $\mathcal{F}$ is an FFT calculation of bin f. The periodogram averaging reduces the variance. There is a trade off between variance reduction and resolution. Experiments have shown that a good compromise is to use a resolution of $M = 32$ frequency bins, thus 5 averages. We are only interested in the magnitude spectrum $P_{x,M}(f,i) = \sqrt{\hat{R}_{x,M}(f,i)}$. The noisy speech spectrum is averaged only over the present block, $\mathbf{x}_L(i)$, using equation (3). The background noise magnitude spectrum can be estimated over a longer time frame during speech pauses by

$$\bar{P}_{w,M}(f,i) = \begin{cases} \mu \bar{P}_{w,M}(f,i-1) + (1-\mu) P_{x,M}(f,i), & \text{noise} \\ \bar{P}_{w,M}(f,i-1), & \text{speech} \end{cases} \tag{4}$$

where μ is the exponential averaging time constant. The speech pauses are detected by a Voice Activity Detector (VAD).

The gain function is formed as

$$G_M(f,i) = \left(1 - k \cdot \frac{\bar{P}^a_{w,M}(f,i)}{P^a_{x,M}(f,i)} \right)^{\frac{1}{a}} \tag{5}$$

where k is the subtraction factor and a controls whether magnitude or power spectral subtraction is used. The combination of k and a controls the amount of noise reduction.

To further reduce the variability of the gain function an adaptive averaging of the gain function, $\bar{G}_M(f,i)$, is employed [2]. The adaptation is controlled by a spectral discrepancy measure between the noise spectra, $\bar{P}_{w,M}(f,i)$, and the noisy speech spectra, $P_{x,M}(f,i)$. This measure is used to adaptively control the averaging time constant. An increased averaging constant is introduced when the magnitude spectrum variations of the input signal are lower, thus giving a reduced variability of the gain function for stationary noise and low energy speech.

The calculated gain function, $\bar{G}_M(f,i)$, is a positive real function which corresponds to a non-causal impulse response. In order to facilitate a short delay causal filter a minimum phase is imposed on the calculated zero phase gain function resulting in the frequency function, $\tilde{G}_M(f,i)$. This phase function is found by using a Hilbert transform relation, see [4]. To obtain the time variable noise reduction filter, $\tilde{\mathbf{g}}_M(i)$, the gain function, $\tilde{G}_M(f,i)$, is transformed to the time-domain. The filtered output signal is obtained by a linear combination of the filter weights, $\tilde{\mathbf{g}}_M(i)$, and input data, $x(n)$, such that

$$y(n) = \sum_{m=0}^{M-1} x(n-m) \cdot \tilde{g}_{M,m}(i). \tag{6}$$

When the shortest delay is chosen the parameter Δ equals 0. Thus, the block of data used to calculate $\tilde{\mathbf{g}}_M(i)$ is at least one block old. For situations where it is possible to increase the delay, the voice quality will improve.

3 Results

Experimental results are presented for noise suppression applied to recorded microphone signals. The speech signals and the background noise signals are recorded separately to facilitate an accurate evaluation. The background noise consists of a recording of computer fan noise. The speech signal is a male voice. The signals are superimposed to yield an input signal of a typical SNR for office handsfree telephony. The recordings have a sampling frequency of 8 kHz and are telephone bandwidth limited.

The resulting noise reduction is controlled by the parameter choices. Using $k = 1$, $a = 1$, $L = 160$ and $M = 32$ yields a 10–14 dB noise reduction while maintaining the speech quality, where the higher amount of noise reduction is obtained for higher SNR input signals.

3.1 Noise Suppression

Since the filter applied is a linear filter, although time-varying between blocks, the speech signal and background noise signal can be filtered separately for evaluation purposes. We then use a pre-calculated filter where the filter calculations have been performed on the combined input data. The sum of the outputs will result in the same output signal as if the noisy speech were filtered, so that

$$\begin{aligned} y(n) = g(x(n), n) = g(s(n) + w(n), n) = \\ g(s(n), n) + g(w(n), n) = y_s(n) + y_w(n) \end{aligned} \tag{7}$$

where $y_s(n)$ and $y_w(n)$ are the processed speech signal and background noise signal, respectively, and $g(\bullet, n)$ denotes a linear time-varying operator. It is possible now to calculate the block energy for the input speech signal, $p_s(i)$, input background noise signal, $p_w(i)$, processed speech signal, $p_{y_s}(i)$, and processed background noise signal, $p_{y_w}(i)$.

To evaluate performance, block-wise SNR and noise reduction measures are defined as

$$\Delta_{\mathrm{SNR}}(i) = \frac{p_w(i)}{p_{y_w}(i)} \cdot \frac{p_{y_s}(i)}{p_s(i)} = \mathrm{SNR}_{out}(i)[\mathrm{dB}] - \mathrm{SNR}_{in}(i)[\mathrm{dB}] \tag{8}$$

which is the signal to noise ratio improvement for block i. The signal to noise ratio before the spectral subtraction is defined as

$$\mathrm{SNR}_{in}(i) = \frac{p_s(i)}{p_w(i)}, \tag{9}$$

and the signal to noise ratio after processing is defined as

$$\mathrm{SNR}_{out}(i) = \frac{p_{y_s}(i)}{p_{y_w}(i)}. \tag{10}$$

The SNR measures are only valid during speech periods. Finally, the noise reduction is defined as

$$\mathrm{NR}(i) = \frac{p_w(i)}{p_{y_w}(i)}. \tag{11}$$

A histogram of the percentage of blocks achieving a certain noise reduction during noise only periods is presented in Figure 2. As seen the noise reduction during noise periods is approximately 14 dB. During speech periods the achieved SNR improvement is dependent on $\mathrm{SNR}_{in}(i)$. For blocks with higher $\mathrm{SNR}_{in}(i)$ it is more difficult to achieve a noise reduction since the gain

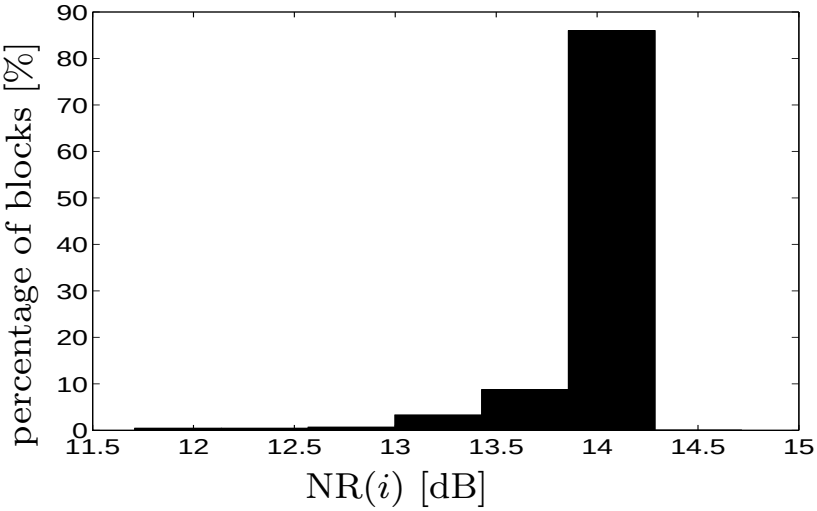

Figure 2: Histogram showing percentage of blocks achieving a certain noise reduction during noise only periods.

function will be close to 1. Noise reduction is, however, not as crucial in periods where the speech signal clearly dominates, because the speech masks the remaining background noise. In Figure 3, a 2-dimensional histogram of the percentage of blocks achieving a certain SNR improvement, $\Delta_{\mathrm{SNR}}(i)$, with a certain $\mathrm{SNR}_{in}(i)$ during speech periods is presented. The $\Delta_{\mathrm{SNR}}(i)$ is in the vicinity of 2 dB during speech periods.

In Figure 4, spectrograms of the input signals 4(a)–4(c) as well as the processed signals 4(d)–4(f) are shown. When comparing the noise only spectrograms 4(c) and 4(f) it can be observed that the noise is less suppressed during the speech periods. In fact the noise is modulated with the noise reduction filter, forming the noise so that it sounds like the speech signal. This modulated noise is mainly masked by the speech but could be perceived as a speech distortion. No sound level changes in the background noise can be perceived between the speech and the noise only periods, although the actual sound levels differ.

3.2 Filtering Delay

The causal noise reduction filter will delay the signal. When imposing a minimum phase filter the delay will be different for each frequency band. We have characterised this delay by means of the group delay. The group delay measures the delay of the envelope of a narrow band signal. Therefore, the influence of each block's filter, $\tilde{\mathbf{g}}_M(i)$, on the output signal, $y(n)$, is presented in Figure 5 as a histogram of the group delay for each frequency band. There

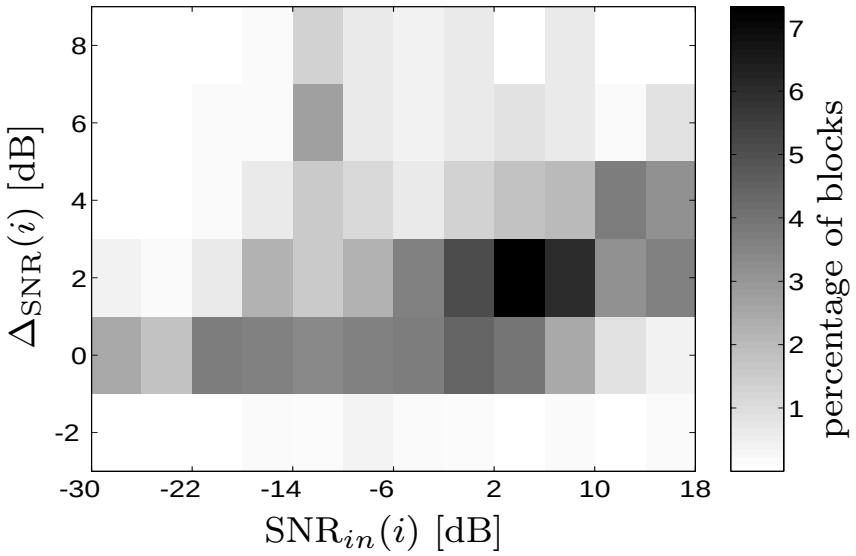

Figure 3: 2D-histogram showing percentage of blocks with certain $\mathrm{SNR}_{in}(i)$ achieving a certain $\Delta_{\mathrm{SNR}}(i)$ during speech periods.

is an energy threshold which is set to 90 dB relative to the frequency having the highest energy component over the block of data in Figure 4(c). Thus it is only the delay of frequency bands containing the speech signal which are evaluated. The observed delay is in the range of 0–7 samples. A causal minimum phase filter can have a negative group delay in the stop-bands which can be observed by noting that a few percent of the frequency bands have a negative delay.

4 Summary and conclusions

The proposed noise reduction method relies on an improved spectral subtraction method. In order to introduce as short a delay as possible the filtering is performed in the time-domain neglecting the fact that the samples should be delayed to achieve the same result as if filtered in the frequency-domain. A minimum phase is imposed on the filter to make it causal with a minimal filtering delay. The result is a good sound quality noise reduction with an SNR improvement of 10–14 dB and a maximum delay of 7 samples. The delay is well within the range of allowed delays in telephone systems. Noise reduction processing is therefore applicable in a wider range of applications, for example conference telephony and office handsfree.

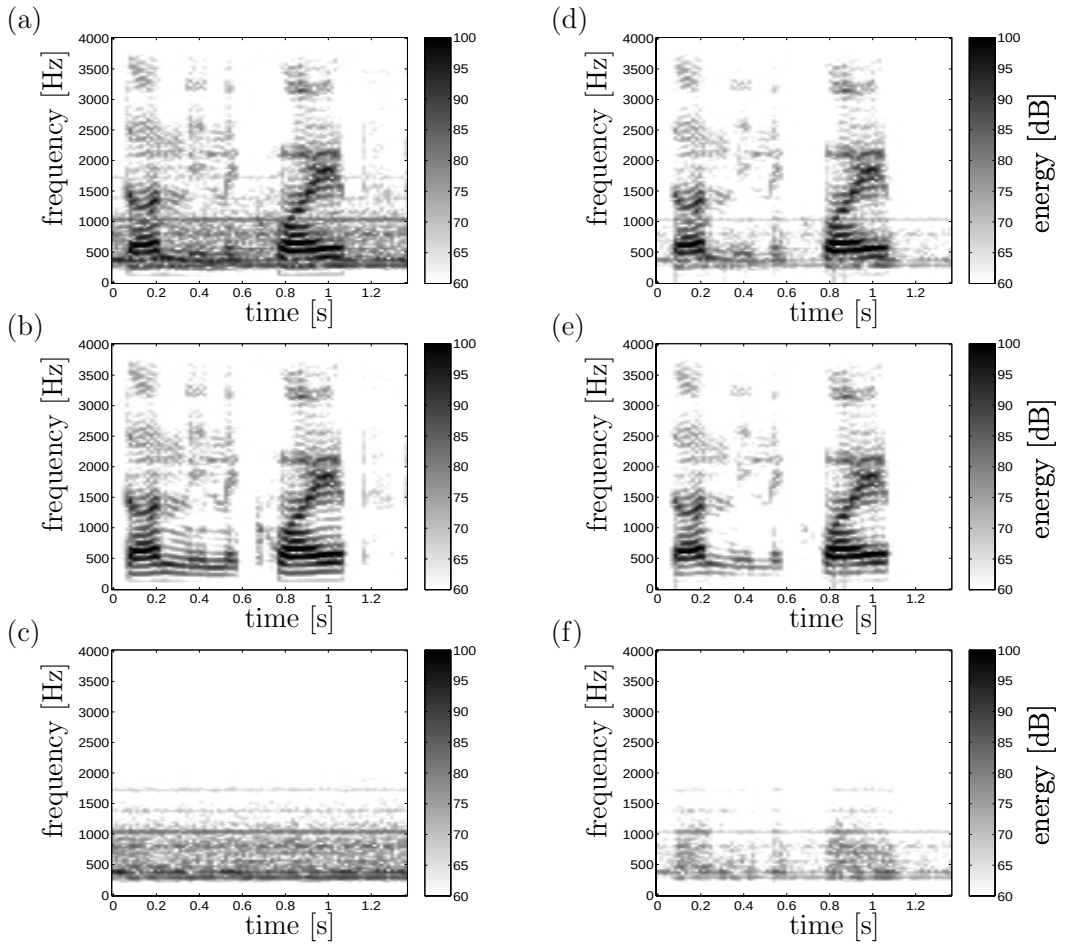

Figure 4: Spectrogram of speech in background fan noise. (a) input noisy speech signal $x(n)$; (b) input speech signal $s(n)$; (c) input noise signal $w(n)$; (d) noise reduced total signal $y(n)$; (e) noise reduced speech signal $y_s(n)$; (f) noise reduced noise signal $y_w(n)$.

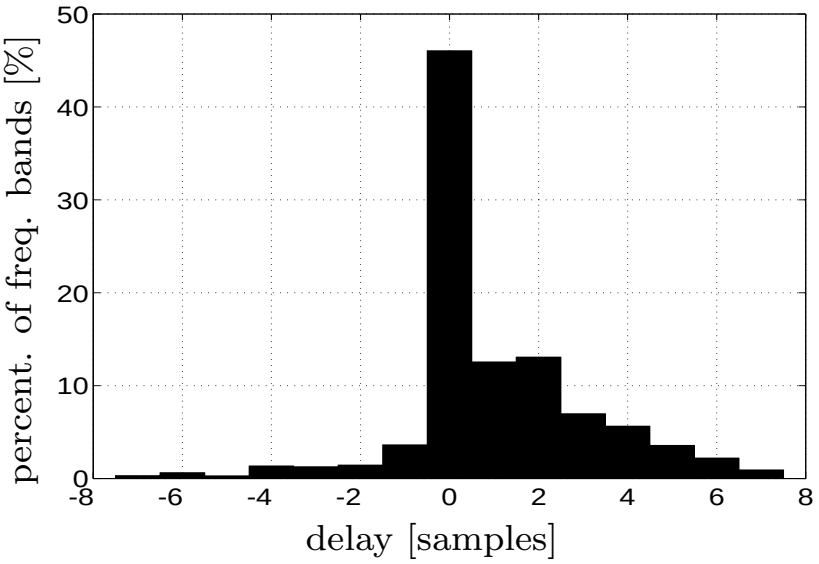

Figure 5: Histogram showing percentage of frequency bands which have a corresponding delay.

5 Acknowledgment

This work was supported by The Foundation for Knowledge and Competence Development and by Ericsson Mobile Communications AB, Sweden.

References

[1] H. Gustafsson, S. Nordholm, I. Claesson: "Spectral Subtraction Using Correct Convolution", *Proc. of ICSV6*, Copenhagen, July 1999.

[2] H. Gustafsson, S. Nordholm, I. Claesson: "Spectral subtraction with adaptive averaging of the gain function", *Proc. of EuroSpeech '99*, Budapest, September 1999.

[3] S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction.", *IEEE Trans. Acoust. Speech and Sig. Proc.*, 27:113-120, 1979.

[4] A. V. Oppenheim, R. W. Schafer: "Discrete - time signal processing.", *Prentice-Hall*, Inter. Ed., 1989.