Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging

Harald Gustafsson, *Member, IEEE,* Sven Erik Nordholm, *Member, IEEE,* and Ingvar Claesson

***Abstract*—In hands-free speech communication, the signal-to-noise ratio (SNR) 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 fast Fourier transform (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 results show a 10-dB background noise reduction for all input SNR situations tested in the range −6 to 16 dB, as well as improvement in speech quality and reduction of noise artifacts as compared with conventional spectral subtraction methods.**

***Index Terms*—Delay effects, mobile communications, signal processing, spectral domain analysis, speech enhancement, telephone sets, vehicles.**

I. 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 hands-free accessories in order to have both hands available and focus better on the driving. When using hands-free accessories, the microphone is situated as far as 1 m 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 toward 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. The musical tones are mainly due to variance in the spectrum estimates [2]. 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 [3], or to hide the musical tones phenomenon using the masking properties of the auditory system [4], [5]. Another known method to reduce the musical tones is to use an over-subtraction factor (larger than one) in combination with a spectral floor [6]. This method has the disadvantage of degrading the speech when musical tones are sufficiently reduced. Ephraim and Malah have suggested a minimum mean-square error estimator of the short-time amplitude spectrum [7]–[9]. This estimator principally results in a low variance estimate of the spectra and hence a reduction in musical tones.

The suggested method deals with four different issues:

- variability of the gain function;
- block effects from the fast Fourier transform (FFT)- and inverse fast Fourier transform (IFFT)-operations;
- noncausal filtering due to the zero-phase gain function;
- short delay processing.

The variance of the gain function is reduced by dividing the current input block into subblocks 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 FFTs, the block effects must be considered. The FFT corre-

Manuscript received June 21, 1999; revised July 19, 2001. This work was supported by the Foundation for Knowledge and Competence Development and by Ericsson Mobile Communications AB, Sweden. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dirk van Compernolle.

H. Gustafsson is with ITS, Blekinge Institute of Technology, 372 25 Ronneby, Sweden and also with Ericsson Mobile Communications, 221 83 Lund, Sweden (e-mail: hgu@its.bth.se).

S. Nordholm is with ATRI, Curtin University of Technology, Perth, WA 6102, Australia.

I. Claesson is with ITS, Blekinge Institute of Technology, 372 25 Ronneby, Sweden.

Publisher Item Identifier S 1063-6676(01)09672-9.

sponds to a critical sampled filter bank [10]. 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% 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 noncausal 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. The idea of replacing the circular convolution by a linear convolution for spectral subtraction is not new [11]. Furthermore, by imposing a linear or minimum phase constraint the gain function will correspond to a causal filter. The conventional gain function in spectral subtraction, which is only an amplitude function, corresponds to a noncausal 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. Many telecommunication systems specify a total system maximum delay for a speech signal. For a GSM system, see [12]. The delay each operation demands is taken from the maximum total delay resource. Choosing a low-delay noise reduction method makes it possible to introduce other features which would otherwise cause an unacceptable maximum delay.

The suggested methods are presented in Section II. This section also includes a short introduction to conventional spectral subtraction. Further, in Section III, the experimental results are presented. Section IV concludes the paper.

II. 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

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

where $\mathbf{x}_N(i)$ is a vector containing the ith 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

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

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 FFTs 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 noncausal 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 noncausal filtering leads to discontinuities between blocks, thereby giving rise to inferior speech quality. The gain function is given by

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

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

A. 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 (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, $\overline{P}_{w,M}(f, i)$, 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 $M \ll N$

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

The shorter periodogram estimates are computed by using subblocks of the input frame, $\mathbf{x}_L(i)$, combining a Bartlett method [13], which is used to decrease the variance of the estimated spectrum, and the rule according to (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, noncausal 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 [13], is achieved by imposing a linear phase on the gain function. This results in a filter complying with

$$g_M(n) = \pm g_M(M - 1 - n) \tag{9}$$

with a delay of $(M - 1)/2$ samples, i.e., a noninteger sample delay since M is even, implied by the use of FFTs.

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 [14]. 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.

In general, many filter design methods can be used to calculate a corresponding phase function on the basis of the amplitude function [15]. For implementation purposes we are limited to low computational complexity methods, but when the computational resources are sufficiently large, a linear or quadratic programming filter design method can be used [16], [17].

B. Further Variability 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

$$\overline{G}_{M,1}(f, i) = \alpha_1(i) \cdot \overline{G}_{M,1}(f, i - 1) + (1 - \alpha_1(i))G_M(f, i) \tag{10}$$

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

$$\alpha_1(i) = 1 - \beta(i). \tag{11}$$

The exponential averaging time in frames is approximately

$$\tau_1(i) \approx \frac{1}{1 - \alpha_1(i)}. \tag{12}$$

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, $\overline{P}_{w,M}(f, i)$

$$\beta(i) = \min\left\{ \frac{\sum_{f=0}^{M-1} \left| \hat{P}_{x,M}(f, i) - \overline{P}_{w,M}(f, i) \right|}{\sum_{f=0}^{M-1} \overline{P}_{w,M}(f, i)}, 1 \right\}. \tag{13}$$

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.

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, $\overline{G}_{M,1}(f, i)$, 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

$$\alpha_2(i) = \begin{cases} \gamma_c\alpha_2(i - 1) + (1 - \gamma_c)\alpha_1(i), & \alpha_2(i - 1) < \alpha_1(i) \\ \alpha_1(i), & \text{otherwise.} \end{cases} \tag{14}$$

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

$$\overline{G}_{M,2}(f, i) = \alpha_2(i) \cdot \overline{G}_{M,2}(f, i - 1) + (1 - \alpha_2(i))G_M(f, i) \tag{15}$$

having an approximate effective averaging time

$$\tau_2(i) \approx \frac{1}{1 - \alpha_2(i)}. \tag{16}$$

To conclude, the adaptive averaging time constant $\alpha_2(i)$ can decrease rapidly but increases only slowly. This averaging scheme forces a slow increase in the averaging time $\tau_2(i)$; the gain function can thus quickly adapt to the new input signal and reduce the "shadow voices."

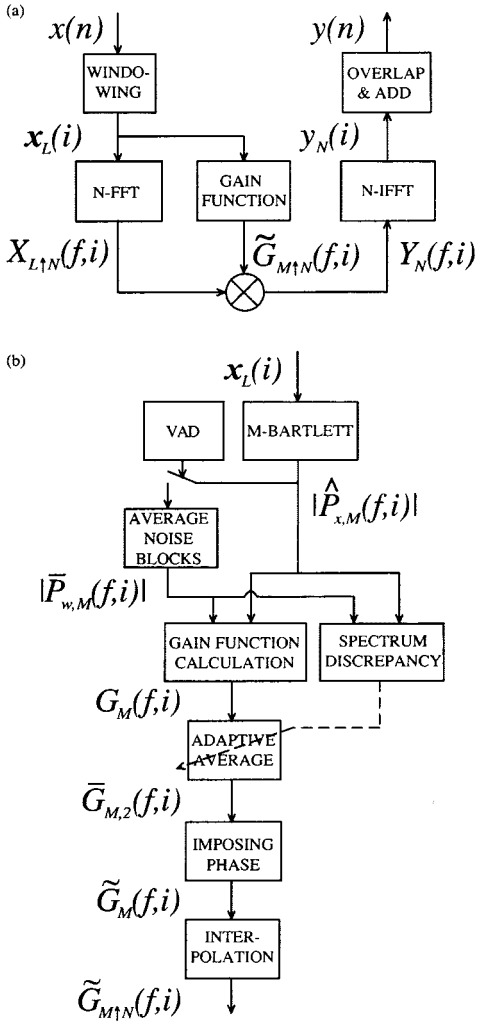

Fig. 1. (a) Outline of the improved spectral subtraction algorithm. (b) Detailed view of the gain function calculation in (a), consisting of the parts facilitating the new causal truly linear filtering and adaptive exponential averaging.

III. 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 [18] 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 http://www.its.bth.se/research.

A. Parameter Choices

The parameter choices derived in this section are mainly directed toward 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 have 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. Choosing a smaller value for the periodogram length M will reduce the variance but results in insufficient frequency resolution. A larger M may be chosen when the delay is of less importance, although an increase in musical tones is to be expected. With larger M values, an increase in speech quality may be achieved due to the larger frequency resolution, particularly when the amplitude spectrum has low variability between frames. Although it is possible to increase the frame length L and the FFT length N, the condition of short-time stationarity sets an upper limit for the frame length of about 40 ms for speech signals. The fact that the frame length L does not affect the filtering delay, and does not need to be a power of two, can be utilized to match the frame length to other system-specific buffer sizes.

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."

B. Degree of Noise Reduction

There is a tradeoff 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 Figs. 2 and 3, respectively. The combined input signal is presented in Fig. 4. The noise reduced output is illustrated in Fig. 5.

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 precalculated 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

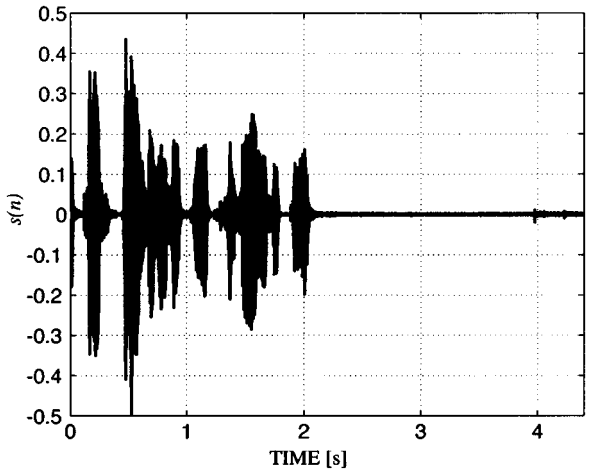

Fig. 2. Input speech signal.

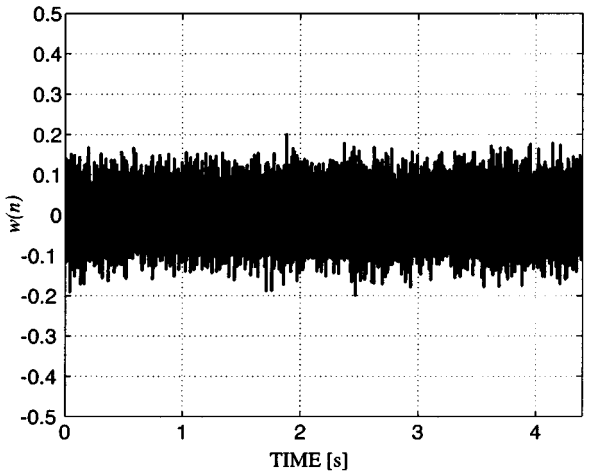

Fig. 3. Input background noise signal.

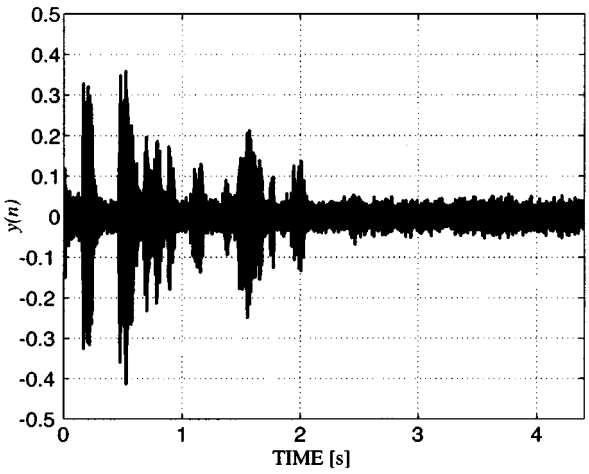

Fig. 5. Output speech when noise reduction is employed.

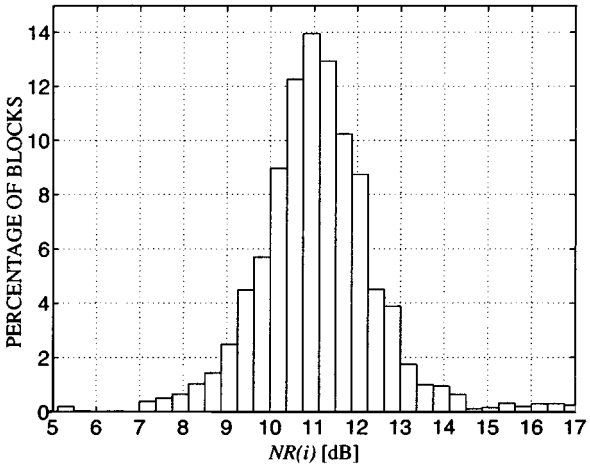

Fig. 6. Histogram over percentage of blocks with a certain noise reduction during noise only periods.

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

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

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 $\mathrm{SNR}_{out}(i)$

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

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

Finally, the noise reduction, $\mathrm{NR}(i)$, is defined as

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

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 Fig. 6. 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 $\mathrm{SNR}_{in}(i)$. Fig. 7 presents a two-dimensional (2-D) 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 one. Noise reduction is, however, not as crucial in periods where the speech signal clearly dominates, due to masking effects. In Fig. 8, a 2-D 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 for all input SNR situations tested in the range -6 to 16 dB.

C. Avoiding Undesirable Periodicity Effects

The experimental results presented in this section show the importance of having the appropriate impulse response length

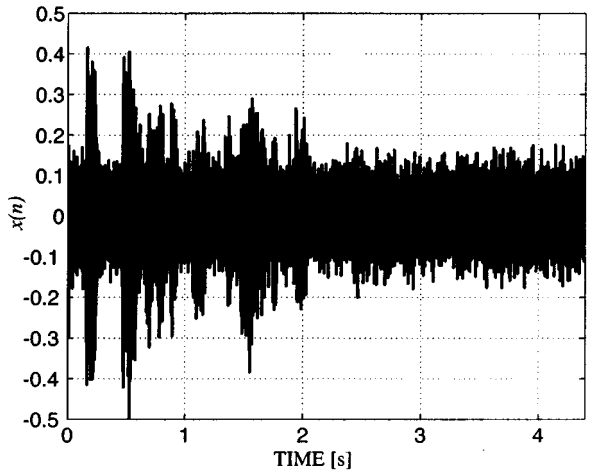

Fig. 4. Input noisy speech signal.

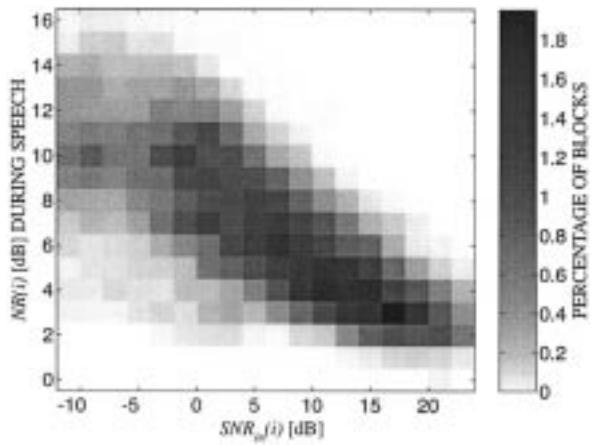

Fig. 7. Two-dimensional histogram over percentage of blocks with certain noise reduction, $\mathrm{SNR}_{in}(i)$, and achieving a certain $\mathrm{NR}(i)$ during speech periods.

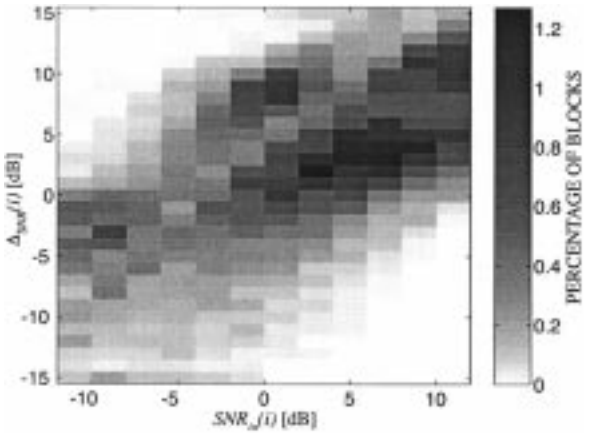

Fig. 8. Two-dimensional histogram over percentage of blocks with certain $\mathrm{SNR}_{in}(i)$ achieving a certain $\Delta_{\mathrm{SNR}}(i)$ during speech periods.

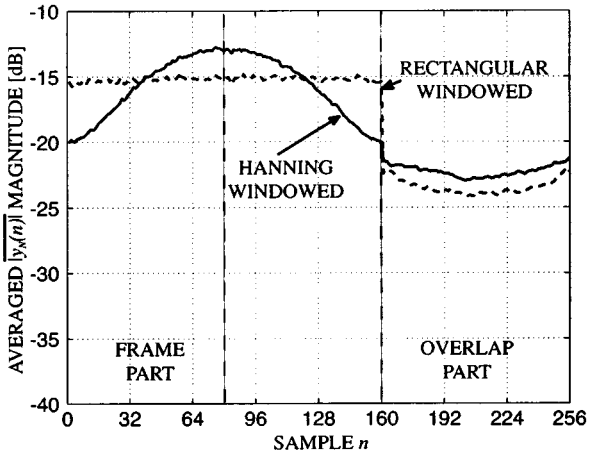

Fig. 9. Mean absolute value output block, $\overline{|y_N(n)|}$, when using conventional spectral subtraction, i.e., filtering with periodic convolution and a zero phase filter. The input time block has been windowed by a rectangular or Hanning window. The vertical lines show the borders for overlap 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.

of the gain function as well as causal properties. 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 Figs. 9 and 10 are absolute mean averages of the output from the IFFT, $\overline{|y_N(n)|}$ (see Fig. 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., noncausal 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% 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, $(M-1)/2$ 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.

D. 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 advan-

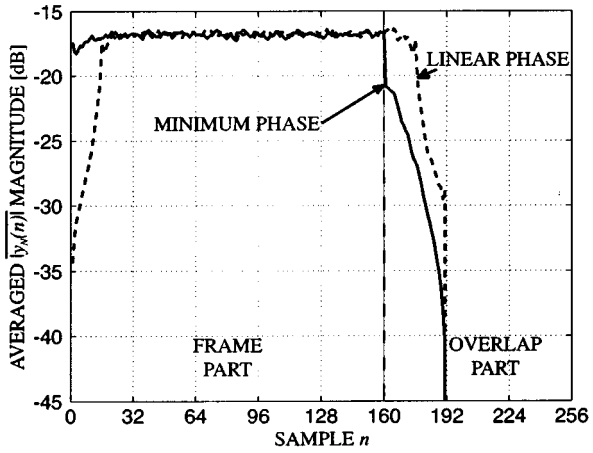

Fig. 10. Mean absolute value output block, $\overline{|y_N(n)|}$, when using the proposed spectral subtraction, i.e., filtering with truly linear convolution and a linear phase or minimum phase filter. The vertical line shows the border for overlap between frames. The phase applied to the gain function makes this causal. The benefit with causal filters can be observed by the low level of the samples in the overlap part.

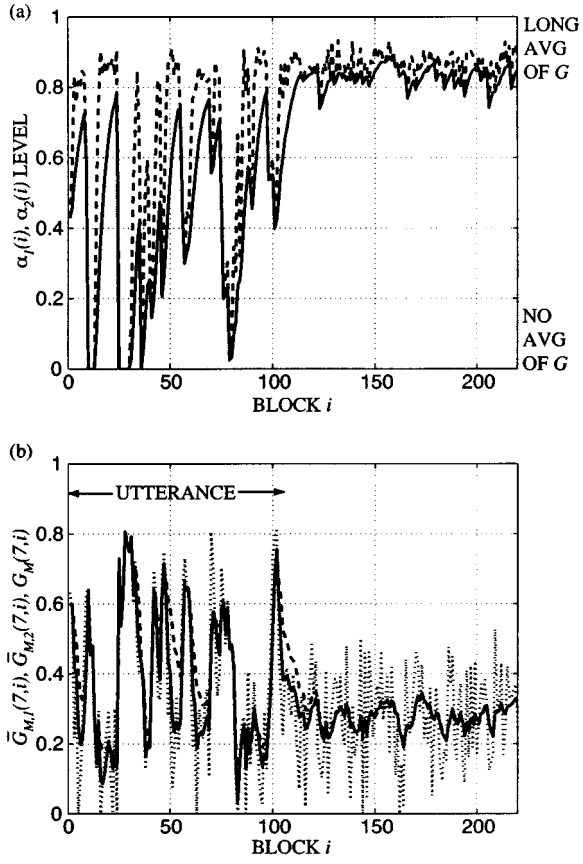

Fig. 11. (a) Averaging time parameter versus block i. Solid line: $\alpha_2(i)$; dashed line: $\alpha_1(i)$. (b) Frequency bin 7 of the gain function versus block i. Solid line: $\overline{G}_{M,2}(7,\, i)$; dashed line: $\overline{G}_{M,1}(7,\, i)$; dotted line: using zero averaging of the gain function $G_M(7,\, i)$. The parameter $\gamma_c = 0.8$.

tage is thereby the reduction of residual noise, i.e., musical tones in the background. Fig. 11 illustrates the behavior of the averaging time parameters and how these affect the gain function. First, the figures show the operation during an utterance where the averaging time is changed rapidly to handle speech with typical small pauses between words. Toward the end the averaging

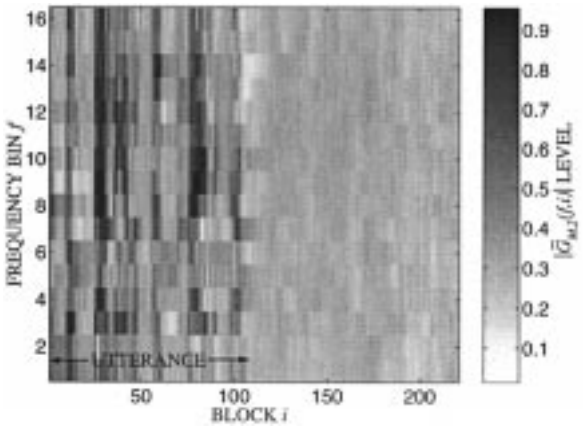

Fig. 12. Absolute value of the gain function, $|\overline{G}_M(f,\, i)|$, with exponential averaging "on."

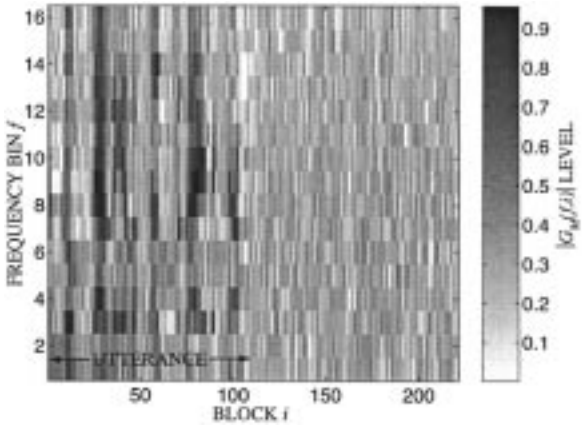

Fig. 13. Absolute value of the gain function, $|G_M(f,\, i)|$, with exponential averaging "off."

time is at a high level since only background noise with slow variations is present. The full gain function with and without exponential averaging is presented in Figs. 12 and 13, respectively. As can be seen in the figures, the variability of the gain function 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.

E. 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, [19], [20]. 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 the Appendix. The objective quality measures mainly indicate the speech quality improvement resulting from the adaptive averaging method since they are too coarse to register discontinuities between blocks. 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

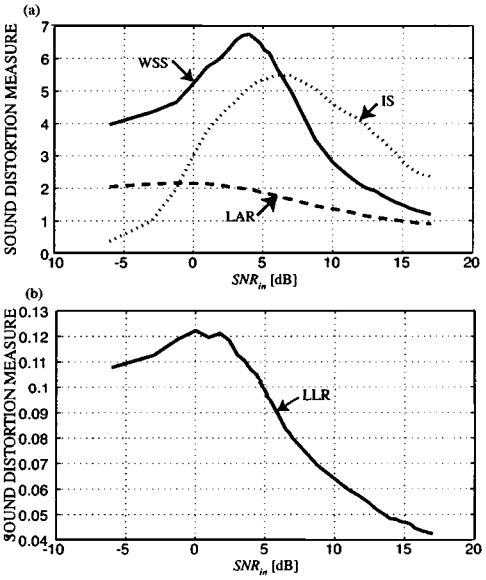

Fig. 14. Objective sound distortion measures for the processed speech, $y_s(n)$. (a) WSS, IS, and LAR measures and (b) LLR measure. The distortion measures are close to zero for clean speech and higher values indicate a greater distortion.

TABLE I
OBJECTIVE SOUND QUALITY MEASURES FOR SPECTRAL SUBTRACTION METHODS DURING SPEECH PERIODS

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
SS WITH FLOOR	1.53	0.47	4.19	67.0
PRESENTED SS	0.32	0.15	2.44	41.5

measures are near zero, and the more distortion in the signal the higher the distortion measure, as expected. In Fig. 14, the objective sound distortion measures are plotted, showing a peak in distortion for total input SNRs 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 I 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 [6], 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 II). The objective quality measures indicate good sound quality and informal listening tests confirm these results. Informal listening tests also confirm that the causal filtering removes the block discontinuities.

IV. CONCLUSIONS AND 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.

TABLE II
OBJECTIVE SOUND QUALITY MEASURES FOR SPECTRAL SUBTRACTION METHODS DURING NOISE ONLY PERIODS

METHOD	IS	LLR	LAR	WSS
CONVENTIONAL SS	75.4	0.23	2.90	34.0
SS WITH FLOOR	279	0.18	2.95	45.0
PRESENTED SS	192	0.06	1.84	4.30

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. The sound quality is further enhanced by adding causal properties to the gain function instead of using a conventional Hanning window to reduce discontinuities between blocks.

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 combined using the overlap-add method which gives improved sound quality even when no Hanning window is applied. 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 extension 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 [21].

APPENDIX
OBJECTIVE SOUND QUALITY MEASURES

The IS distortion measure is derived from the 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

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

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, but by taking the logarithm

of LLR the same equilibrium can be achieved at the price of a somewhat different scale.

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.

LAR measure is calculated by

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

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

The 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 valley. The weighting function in the distortion measure put emphasize on formant placement. For a more precise definition, see [20]. An advantage of this measure is the aural model which gives a spectral distance incorporating perceptually meaningful frequency weighting.

ACKNOWLEDGMENT

The authors would like to thank J. Rasmusson, at Ericsson Mobile Communications, for his support of this research. They would also like to thank T. Samuels for proofreading the manuscript.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, 1979.
[2] P. Händel, "Low-distortion spectral subtraction for speech enhancement," in *Proc. Eurospeech '95*, vol. 2, Sept. 1995, pp. 1549–1552.
[3] M. Winberg and I. Claesson, "Spectral subtraction with extended methods,", Res. Rep. HK-R, Aug. 1996.
[4] N. Virage, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.
[5] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria," in *Proc. IEEE ICASSP*, vol. 2, 1993, pp. 359–362.
[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, 1979, pp. 208–211.
[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE ASSP*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[8] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE ASSP*, vol. 33, pp. 443–445, April 1985.
[9] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.
[10] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[11] J. Hakkinen and M. Vaananen, "Background noise suppressor for a car hands-free microphone," in *Proc. ICSPAT*, 1993, pp. 300–307.
[12] European Telecommunications Standards Institute, "Digital cellular telecommunications system (Phase 2+)," *Transmission Planning Aspects of the Speech Service in the GSM Public Land Mobile Network (PLMN) (GSM 03.50)*, 2000.
[13] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing; Principles, Algorithms, and Applications*, 2nd ed. New York: Macmillan, 1992.
[14] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[15] T. W. Parks and C. S. Burrus, *Digital Filter Design*. New York: Wiley, 1987.
[16] M. Dahl, S. Nordebo, and I. Claesson, "Complex approximation by semi-infinite quadratic programming," in *Proc. ISPACS*, 2000.
[17] S. Nordebo, M. Dahl, and I. Claesson, "Complex Chebyshev approximation using conventional linear programming," in *Proc. ISPACS*, 2000.
[18] European Telecommunications Standards Institute, "European digital cellular telecommunications system (Phase 2)," *Voice Activity Detection (VAD) (GSM 06.32)*, 1994.
[19] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP 98*, Sydney, Australia, Dec. 1998.
[20] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[21] B. C. J. Moore, Ed., *Hearing*. New York: Academic, 1995.

Harald Gustafsson (M'98) was born in Sweden in 1972. He received the M.S.E.E. degree from Lund University, Lund, Sweden, in 1995, and the Tech. Lic. degree from Blekinge Institute of Technology, Ronneby, Sweden, in 2000, where he is currently pursuing the Ph.D. degree.

Since 1996, he has been a Development/Staff Engineer with Ericsson Mobile Communications, Lund. His current research interests are in speech enhancement and speech bandwidth extension.

Sven Erik Nordholm (M'91) received the Dipl. Eng., Tech. Lic., and Ph.D. degrees from Lund University, Lund, Sweden, in 1983, 1989, and 1992, respectively.

From 1983 to 1986, he was a Development Engineer with GAMBRO, Lund. From 1986 to 1990, he was a Teaching and Research Assistant at Lund University. From 1990 to 1999, he held various positions at Blekinge Institute of Technology, Ronneby, Sweden. Since 1999, he has been a Professor and Director of Australian Telecommunication Research Institute, Perth, Australia. His current research interests are in speech enhancement, echo cancellation and wireless communication.

Ingvar Claesson was born in Sweden in 1957. He received the M.S. degree in 1980 and the Ph.D. degree in 1986 in electrical engineering from the University of Lund, Lund, Sweden.

Since 1990, he has been building a Telecommunication and Signal Processing Department at the new Blekinge Institute of Technology, Ronneby, Sweden, where he is Head of Research and holds the Chair in applied signal processing. His interests are in acoustic signal processing, filter design, adaptive filtering, and noise cancellation.