Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 2001-93

A multichannel Subband GSVD Approach to Speech Enhancement [1]

Ann Spriet[2], Marc Moonen[3] ,Jan Wouters[4]

Published in European Transactions on Telecommunications, Special Issue on Acoustic Echo and Noise Control, Vol. 13, No. 2, pp. 149-158, March/April 2002

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/spriet/reports/01-93.ps.gz*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, Tel. 32/16/32 18 99, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *ann.spriet@esat.kuleuven.ac.be*. K.U.Leuven, Lab. Exp. ORL, Dept. Neurowetenschappen, Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 24 15, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm*. Ann Spriet is a Research Assistant supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) - Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the Katholieke Universiteit Leuven, in the framework of the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government, IUAP P4-02 (1997-2001) 'Modeling, Identification, Simulation and Control of Complex Systems' and FWO Research Project nr. G.0233.1 ('Signal processing and automatic patient fitting for advanced auditory prostheses'). The scientific responsibility is assumed by its authors.

[3]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *marc.moonen@esat.kuleuven.ac.be*. Marc Moonen is a professor at the Katholieke Universiteit Leuven.

[4]K.U.Leuven, Lab. Exp. ORL, Dept. Neurowetenschappen, Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 23 42, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm* E-mail: *jan.wouters@uz.kuleuven.ac.be*. Jan Wouters is a professor at the Katholieke Universiteit Leuven.

Acoustic Echo and Noise Control

A Multi-Channel Subband Generalized Singular Value Decomposition Approach to Speech Enhancement

ANN SPRIET[1,2], **MARC MOONEN**[1], **JAN WOUTERS**[2]
[1]K.U. Leuven, ESAT-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee
{spriet,moonen}@esat.kuleuven.ac.be
[2]K.U. Leuven, Lab. Exp. ORL, Kapucijnenvoer 33, B-3000 Leuven
jan.wouters@uz.kuleuven.ac.be

Abstract. In speech communication applications, the presence of noise (and reverberation) in the environment may render the captured speech unintelligible. Noise reduction algorithms are thus needed to improve the intelligibility. Recently, a multi-channel noise reduction technique, based on a Generalized Singular Value Decomposition (GSVD) has been proposed. The GSVD based filter minimizes the Mean Square Error (MSE) between the desired signal portion in the received signals and the sum of the filtered received microphone signals. In this paper, we propose a subband implementation of the GSVD based filter. It is shown that - in the case of coloured signals - the subband implementation improves intelligibility more than the fullband approach. In addition, a significant complexity reduction is achieved, allowing for low cost, real-time applications.

1 INTRODUCTION

In speech communication applications, such as teleconferencing, handsfree telephony, hearing aids, the quality of the speech signal recordings is seriously degraded by the presence of background noise and/or reverberation. This stems from the large distance between the speaker and the microphones. Besides the quality, the presence of noise may deteriorate the intelligibility of the speech drastically, making the use of a noise reduction algorithm necessary. Multi-microphone noise reduction techniques exploit both spatial and spectral information to distinguish between the desired and interfering signals and are thus preferred to single microphone procedures (such as spectral subtraction).
The most common adaptive multi-microphone noise reduction technique is the Generalized Sidelobe Canceller (GSC) [1][2]. The GSC assumes the position of the desired source, as well as the microphone characteristics and positions to be known. Deviations from these assumptions are harmful to its performance. In [3], a multi-channel signal enhancement technique based on a Generalized Singular Value Decomposition (GSVD) has been proposed. The GSVD based optimal filter minimizes the Mean Square Error (MSE) between the desired signal portion in the received signals and the sum of the filtered received microphone signals. Since no assumptions are made about the location of the desired speaker, the microphone characteristics and positions, this algorithm is shown to be more robust than the GSC [3][4].

In this paper, we propose a subband implementation of the GSVD based filter. In speech communication applications, the major purpose of a noise reduction algorithm is to improve the intelligibility of the recorded speech. The fullband GSVD procedure minimizes the fullband MSE. Most acoustic interferers have their power concentrated at the low frequencies e.g. car noise, speech interference. In this case, the fullband GSVD weighs the MSE more at the high energy low frequencies than at the low energy high frequencies. This may result in a higher speech distortion and/or smaller amount of noise reduction at the high frequencies. However, all frequencies contribute independently to intelligibility. Hence, an improvement in Signal-to-Noise Ratio (SNR) in any of the frequency bands, especially in the frequency range from 500 Hz to 4000 Hz, can increase intelligibility. Minimizing the MSE independently in different frequency bins may thus improve intelligibility. Section 2 briefly reviews the GSVD based optimal filter. The subband implementation is motivated and described in section 3. Section 4 and section 5 compare the computational cost and the performance of the fullband and subband approach. It is shown that the subband implementation improves intelligibility more than the fullband approach, and furthermore that a significant complexity reduction is achieved.

2 GSVD BASED OPTIMAL FILTER

Consider an array of M microphones and let

$$u_i[k] = h_{si}[k] \otimes s[k] + n_i[k] = d_i[k] + n_i[k] \quad (1)$$

be the k-th signal sample at the i-th microphone, where $d_i[k]$ is the desired speech portion, i.e. the desired source signal $s[k]$ convolved with the acoustic path $h_{si}[k]$ from the source to the i-th microphone, and $n_i[k]$ the environment plus internal noise. Define L as the filter length per channel. Construct the vector $\mathbf{u}_k = [\mathbf{u}_{1,k}^T \quad \mathbf{u}_{2,k}^T \cdots \mathbf{u}_{M,k}^T]^T$ with $\mathbf{u}_{i,k} = [u_i[k] \quad u_i[k-1] \cdots u_i[k-L+1]]^T$, and similarly the vectors $\mathbf{d}_k$ and $\mathbf{n}_k$. The optimal filter $\mathbf{W} \in \mathcal{C}^{ML \times ML}$ produces speech estimates $\hat{\mathbf{d}}_k = \mathbf{y}_k = \mathbf{W}^H \mathbf{u}_k$ such that the MSE $\text{diag}\left\{\varepsilon\{\mathbf{e}_k \mathbf{e}_k^H\}\right\}$ with $\mathbf{e}_k = \mathbf{d}_k - \mathbf{W}^H \mathbf{u}_k$ is minimized, and is given as:

$$\mathbf{W} = \varepsilon\{\mathbf{u}_k \mathbf{u}_k^H\}^{-1} \varepsilon\{\mathbf{u}_k \mathbf{d}_k^H\}. \quad (2)$$

Assuming that the speech and noise signal are uncorrelated, (2) becomes:

$$\mathbf{W} = \varepsilon\{\mathbf{u}_k \mathbf{u}_k^H\}^{-1} (\varepsilon\{\mathbf{u}_k \mathbf{u}_k^H\} - \varepsilon\{\mathbf{n}_k \mathbf{n}_k^H\}). \quad (3)$$

The second order statistics of the noise signal are assumed to be quite stationary such that they can be estimated during periods of noise-only.
The filter $\mathbf{W}$ is approximated at time instant k by means of a GSVD of an input data matrix $\mathbf{U}_k \in \mathcal{C}^{p \times ML}$ and a noise data matrix $\mathbf{N}_k \in \mathcal{C}^{q \times ML}$

$$\begin{array}{rcl} \mathbf{U}_k & = & \left[\begin{array}{cccc} \mathbf{u}_k & \mathbf{u}_{k-1} & \cdots & \mathbf{u}_{k-p+1} \end{array}\right]^T \\ \mathbf{N}_k & = & \left[\begin{array}{cccc} \mathbf{n}_k & \mathbf{n}_{k-1} & \cdots & \mathbf{n}_{k-q+1} \end{array}\right]^T \end{array} ; \quad (4)$$

collecting respectively p speech+noise and q noise-only signal vectors (corresponding to speech pauses) up to time instant k:

$$\left\{\begin{array}{rcl} \mathbf{U}_k & = & \mathbf{Q}_\mathbf{U}.\text{diag}\{\sigma_i\}.\mathbf{X}^H \\ \mathbf{N}_k & = & \mathbf{Q}_\mathbf{N}.\text{diag}\{\eta_i\}.\mathbf{X}^H \end{array}\right. \quad (5)$$

with $\mathbf{Q}_\mathbf{U}, \mathbf{Q}_\mathbf{N}$ orthogonal matrices, $\mathbf{X}$ an invertible matrix and $\frac{\sigma_i}{\eta_i}$ the generalized singular values.
Using time averaging to estimate the 2nd order statistics[1]

$$\begin{array}{rcl} \varepsilon\{\mathbf{u}_k.\mathbf{u}_k^H\} & \simeq & \frac{1}{p}\mathbf{U}_k^T \mathbf{U}_k^* \\ \varepsilon\{\mathbf{n}_k.\mathbf{n}_k^H\} & \simeq & \frac{1}{q}\mathbf{N}_k^T \mathbf{N}_k^* \end{array} \quad (6)$$

and substituting (5) in (2), gives:

$$\mathbf{W}_k^* \simeq \mathbf{X}^{-H} \text{diag}\left\{\max\left(1 - \frac{p}{q}\frac{\eta_i^2}{\sigma_i^2}, 0\right)\right\} \mathbf{X}^H. \quad (7)$$

Since theoretically $\varepsilon\left\{1 - \frac{p}{q}\frac{\eta_i^2}{\sigma_i^2}\right\} \geq 0$, negative values of $1 - \frac{p}{q}\frac{\eta_i^2}{\sigma_i^2}$ are set to 0.

[1] This assumes stationarity and ergodicity of the signals u and n.

Filtering $\mathbf{u}_k$ with the l-th column of $\mathbf{W}_k$ yields the estimate of the l-th element of $\mathbf{d}_k$,

$$\hat{d}_i[k - \Delta_l] = \mathbf{W}_k(:, l)^H \mathbf{u}_k, \quad (8)$$

with $i = \text{mod}(l-1, L) + 1$ and $\Delta_l = \text{rem}(l-1, L)$. The optimal choice for l corresponds to the smallest element on the diagonal of the error covariance matrix $\varepsilon\{\mathbf{e}_k \mathbf{e}_k^H\}$. While it is perfectly possible to construct this matrix and hence define the smallest diagonal element [3], in practice, using the column $l = \left\lceil \frac{L}{2} \right\rceil + kL$ with $k \in [0, M-1]$ is found not to decrease performance. In the rest of this document, column $l = \left\lceil \frac{L}{2} \right\rceil$ will be used. In this way, we obtain an estimate of the speech signal received at the first microphone delayed by $\left\lceil \frac{L}{2} \right\rceil - 1$ samples at the rate of the GSVD. Finally, suffice it to say that, in practice, the GSVD is not calculated from scratch at each time k but updated recursively [3]. To cope with time variations in the position and average power spectral density of the speech and the noise sources, exponential weighting factors λ_s and λ_n are incorporated in the estimation of the second order speech + noise and noise-only statistics, respectively. An exponential weighting factor λ_s or λ_n corresponds to $p \approx \frac{1}{1-\lambda_s^2}$ and $q \approx \frac{1}{1-\lambda_n^2}$, respectively, in (6).

3 SUBBAND GSVD APPROACH

3.1 IMPROVED INTELLIGIBILITY AT LOWER COST

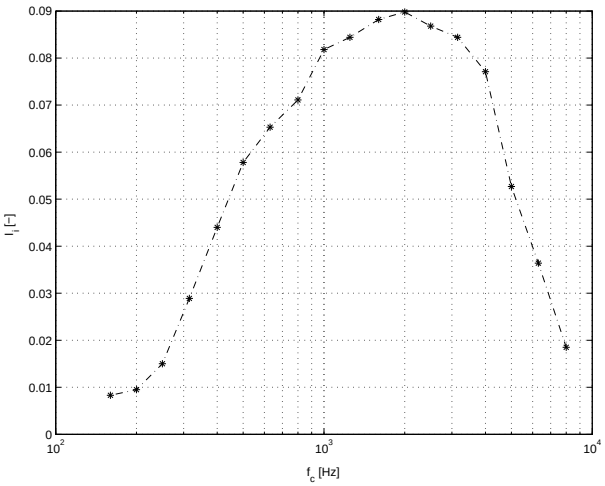


Figure 1: Band importance function I_i as a function of the center frequency f_c of the one-third octave band.

In speech communication applications, the major purpose of a noise reduction algorithm is to improve the intelligibility of the recorded speech signal, rather than to maximize overall SNR or to minimize an MSE criterion. The performance of a noise reduction algorithm is therefore assessed in terms of speech intelligibility weighted measures [5] such as the intelligibility weighted SNR, denoted

as $\text{SNR}_{\text{intellig}}$:

$$\text{SNR}_{\text{intellig}} = \sum_i I_i \text{SNR}_i, \tag{9}$$

where the band importance function I_i expresses the importance of the i-th one-third octave band for intelligibility [6] and SNR_i equals the SNR in dB in the i-th one-third octave band. Figure 1 plots the band importance function I_i, which is used in the sequel to calculate $\text{SNR}_{\text{intellig}}$, as a function of the center frequency f_c (in Hz) of the one-third octave band. These measures, based on the articulation index, reflect the fact that distinct frequency bands contribute independently and differently to intelligibility. They have been introduced to assess the objective performance of noise reduction algorithms for hearing aid applications.

Most acoustic interferers are broadband with their power concentrated at the low frequencies e.g. car noise, speech interference. Hence, their spectrum resembles the average long-term speech spectrum. In this case, minimizing the fullband MSE will weigh the MSE (i.e. sum of distortion and noise) more at the high energy low frequencies than at the low energy high frequencies. If the number of degrees of freedom (and thus the filter length) is small, the filter taps are mainly used to improve the MSE at the low frequencies, resulting in a small improvement and larger distortion of the desired signal at higher frequencies (see also section 5). In [7], a prewhitening operation is proposed to improve the intelligibility obtained with a GSC. Such prewhitening is intrinsically present in a subband implementation. A subband implementation or a prewhitening operation will attenuate the interferer and minimize the distortion of the desired signal over a broader frequency range. Although the bandwidth of the frequency bins contributing equally to intelligibility increases as a function of frequency [6] and hence, the noise reduction at low frequencies contributes more to $\text{SNR}_{\text{intellig}}$, paying more attention to the higher frequencies may yield a significant increase in intelligibility. In some applications, only a limited microphone array size can be used, e.g., in hearing aids or in a car-environment. However, a small-sized array can better reduce correlated noise at the high frequencies than at the low frequencies. In this case, the additional noise reduction offered by the subband approach or the prewhitening at the higher frequencies can significantly increase $\text{SNR}_{\text{intellig}}$. Maximization of intelligibility can thus better be achieved by minimizing the MSE in separate subbands. Since a subband implementation results in a smaller amount of residual noise and distortion at high frequencies, it may also improve the speech quality provided no degradation occurs at the lower frequencies.

In addition, because a subband GSVD can be performed at a lower rate, a reduction in complexity is achieved (see section 4).

3.2 UNIFORM SUBBAND GSVD

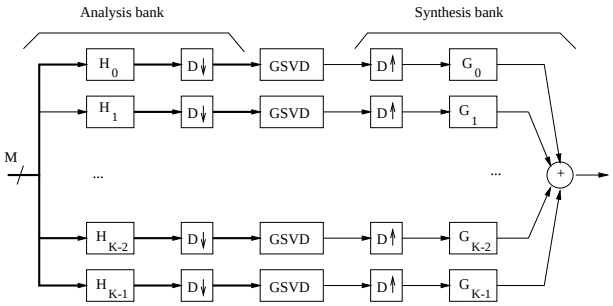


Figure 2: Concept of subband GSVD.

The concept of subband GSVD is illustrated in figure 2. The M microphone signals are each split into K frequency bands by a (nearly) Perfect Reconstruction (PR) (see below), oversampled uniform DFT modulated filterbank [8]. These filterbanks can be implemented efficiently using a polyphase decomposition of the prototype filter $H_0(f)$ and an FFT. Since the fullband signals are real and (roughly) half of the complex DFT modulated analysis filters are the complex conjugate of the other half, only half [2] of the complex subband signals need to be processed.

The optimal GSVD based filter operates on the subband signals at the decimated sampling frequency $\frac{f_s}{D}$. To avoid audible interband aliasing, the downsampling factor D is kept smaller than the number of subbands K. In [9], it is shown that for a system identification setup, the aliasing level in the subbands corresponds to a non-linear distortion and thus presents a lower bound for the minimum achievable MSE in each subband. This indicates that -if the MSE is minimized in each subband independently (and thus no fullband adaptation criterion is used) - the amount of aliasing determines the performance and should be kept small. A large stopband attenuation and small amount of interband aliasing are thus required.

If no filtering (or only a delay) is performed in the subbands, which e.g. occurs in a noise free scenario ($\mathbf{W} = \mathbf{I}$ in (7)), the effect of the analysis and synthesis bank should approximate a delay Δ. This is called a perfect reconstruction (PR) filterbank. The delay Δ caused by a DFT modulated filterbank equals [8]

$$\Delta = D \left\lceil \frac{L_p}{D} \right\rceil + D - 1 \tag{10}$$

samples at f_s, with L_p the length of the prototype filter. In the sequel, a sampling frequency $f_s = 16000$ Hz is used. The delay Δ as well as the cost of the filterbank implementation increases with increasing filter length L_p. In, e.g., hearing aid applications, the total processing delay should

[2] Actually, 2 real and $\frac{K}{2} - 1$ complex signals need to be processed if K is even and 1 real and $\frac{K-1}{2}$ complex signals need to be processed if K is odd. We will approximate this in the complexity calculations later on by $\frac{K}{2}$ complex signals.

be smaller than 20 ms in order not to degrade intelligibility due to asynchronism with lip reading [10]. Similar upper bounds for the processing delay exist for other applications. The total processing delay of the subband GSVD based algorithm consists of the delay Δ of the filterbank and the delay of the GSVD. The latter equals $\Delta_l = rem(l-1, L_{sub})$ samples at $\frac{f_s}{D}$ (see also section 2), with L_{sub} the filter length per channel used in the subbands. Since in the sequel, the column $l = \lceil \frac{L}{2} \rceil$ of $\mathbf{W}_k$ is used in (8), the total processing delay equals

$$\underbrace{D\left\lceil \frac{L_p}{D} \right\rceil + D - 1}_{\Delta} + \underbrace{D\left(\left\lceil \frac{L_{sub}}{2} \right\rceil - 1\right)}_{\text{delay GSVD}} \qquad (11)$$

samples at f_s. For reasonable L_{sub}, the delay of the GSVD is negligible compared to the delay Δ of the filterbank. In order to achieve a large stopband attenuation with reasonable L_p, so-called "nearly PR" filter banks [8] are used. Note that the perfect reconstruction property would be lost anyway when a filtering operation is performed on the subband signals. The deviation from perfect reconstruction is expressed by the reconstruction error [9]

$$RE = \sum_{n=-\infty}^{\infty} (t[n] - \delta[n-\Delta])^2, \qquad (12)$$

where $t[n]$ is the distortion function of the filterbank and Δ the delay caused by the filterbank. Together with the stopband attenuation, the reconstruction error is a measure for the minimum achievable MSE for the fullband signal [9].

In the sequel, two nearly PR oversampled DFT modulated filter banks are used: a 16 subband filter bank with $D = 12$, prototype filter length $L_p = 120$, delay $\Delta = 8.2$ ms, reconstruction error $RE = -34$ dB and a 32 subband filter bank with $D = 20$, $L_p = 240$, $\Delta = 16.2$ ms, $RE = -40$ dB.

4 COMPUTATIONAL COMPLEXITY

4.1 FULLBAND GSVD

Although approximate recursive GSVD algorithms are available [3], the computational complexity of a multichannel fullband GSVD based approach may still be too high for real-time DSP implementations. The cost of a real, recursive square-root free GSVD implementation [3] equals

$$f_s \left(\frac{17.5}{d_g} + \frac{4}{d_f}\right) L^2 M^2 \qquad (13)$$

operations (real multiplications and additions) per second (ops/s). The factors d_g and d_f indicate the number of samples between respectively two consecutive GSVD and filter updates. They trade off convergence speed and cost [3].

4.2 SUBBAND GSVD

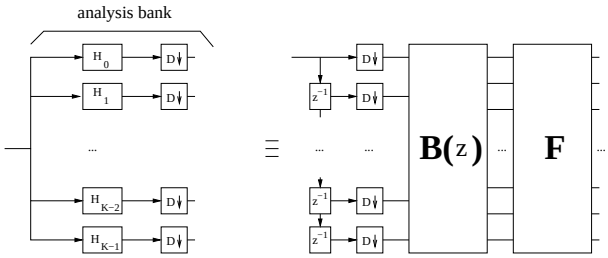


Figure 3: Efficient implementation of a DFT modulated analysis filter bank.

Since in the subband approach, the GSVD operates at a D times lower rate, a significant complexity reduction is achieved. The overall computational cost of the subband GSVD algorithm consists of the complexity of the filter bank and the complexity of the subband GSVDs. When using a DFT modulated filterbank, only half the number of complex subband signals have to be processed. However, a complex recursive GSVD costs about 4 times as much as a real one. The subband GSVDs thus cost about

$$2.K.\frac{f_s}{D}\left(\frac{17.5}{d_g} + \frac{4}{d_f}\right).L_{sub}^2 M^2 \qquad (14)$$

ops/s, where L_{sub} equals the filter length per channel used in the subbands.

A DFT modulated analysis bank can be computed efficiently by factoring the D-th order polyphase matrix $\mathbf{H}(z)$ as $\mathbf{FB}(z)$, with $\mathbf{F}$ a $K \times K$ DFT matrix, where K equals the number of subbands [8]. This is depicted in figure 3. A similar factorisation exists for the synthesis bank. The complexity corresponding to $\mathbf{B}(z)$ equals the complexity of a filter with length L_p (i.e. $= 2L_p$ operations) operated at the rate $\frac{f_s}{D}$. The complexity of a DFT with K a power of 2, requires approximately $2K \log_2 K$ operations. Hence, the implementation of a DFT modulated analysis or synthesis filter bank requires

$$(2L_p + 2K \log_2 K)\frac{f_s}{D} \qquad (15)$$

ops/s. The total complexity of the filter bank implementation (i.e. M analysis banks and 1 synthesis bank) thus equals

$$(M+1)(2L_p + 2K \log_2 K)\frac{f_s}{D} \qquad (16)$$

ops/s. In general, the cost of the filterbank will be negligible compared with the cost of the GSVDs. From (13) and (14), it follows that the subband GSVD will roughly be a factor $\frac{D^3}{2K}$ less complex if the same temporal window or temporal dimension T is used. The temporal window T, expressed as the number of samples at sampling rate f_s, equals the filter length L for the fullband algorithm and $L_{sub}.D$ for the subband approach. This amounts to roughly a factor 54 and 125, respectively, for the above given setups.

4.3 COMPARISON

Table 1 compares the complexity (expressed in Mops/s) of the recursive fullband and subband approach (16 and 32 subbands) with $d_f = d_g = 1$ for different number of microphones M and temporal windows T. The complexity of subband GSVD makes a multi-channel real-time DSP implementation feasible. A further decrease in complexity can be obtained by reducing the number of filter and GSVD updates by a factor d_f and d_g respectively and/or by using only a subset of the microphones in each frequency bin.

Table 1: Complexity (in Mops/s) of the recursive fullband and subband GSVD for different number of microphones M and different temporal windows T, $d_f = d_g = 1$.

		fullband	subband	
			$K = 16$ $D = 12$	$K = 32$ $D = 20$
$M = 2$	$T \approx 24$	792.6	16.2	6.3
	$T \approx 48$	3170.0	60.2	19.5
$M = 3$	$T \approx 24$	1783.3	35.0	12.5
$M = 4$	$T \approx 24$	3170.0	61.2	20.8
	$T \approx 48$	12681.2	237.3	73.7

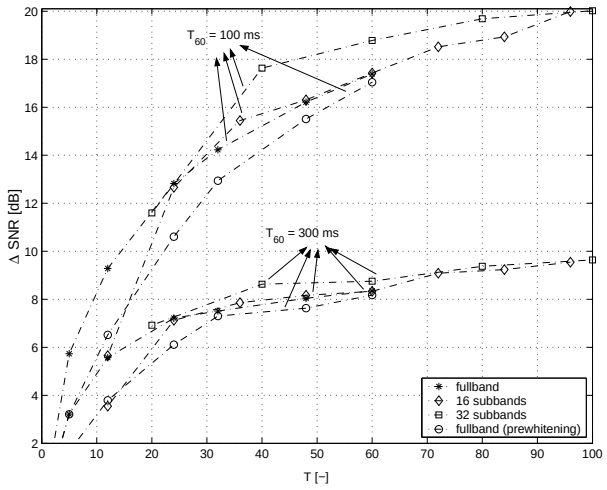


Figure 4: Improvement in SNR, ΔSNR, as a function of temporal window length T for the subband and fullband (with and without prewhitening) GSVD implementation, $d_f = d_g = 1$.

5 SIMULATION RESULTS

5.1 SETUP

The performance of the subband and fullband approach is evaluated based on simulations. A room with dimensions 7 m x 4 m x 2.5 m is simulated using the image method

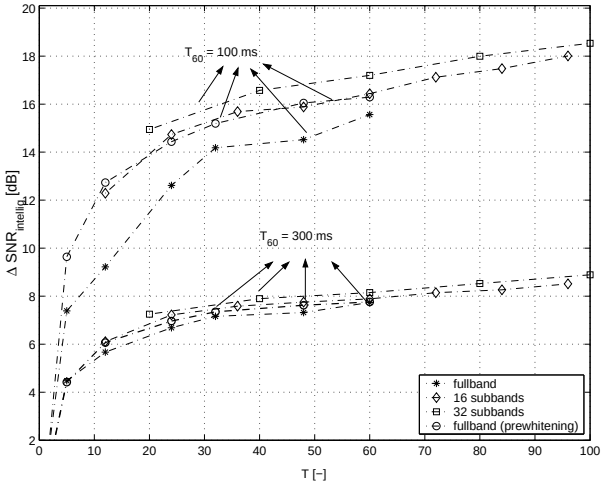


Figure 5: Improvement in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$, as a function of temporal window length T for the subband and fullband (with and without prewhitening) GSVD implementation, $d_f = d_g = 1$.

Table 2: Coordinates of microphones, desired and interfering source.

	x	y	z
dimensions room	7 m	4 m	2.5 m
position mic 1	3 m	1 m	1 m
position mic 2	3.02 m	1 m	1 m
position mic 3	3.06 m	1 m	1 m
position mic 4	3.14 m	1 m	1 m
position speaker	2.4 m	2 m	1.5 m
position interferer	4 m	3 m	1 m

[11]. A 4-microphone array with logarithmic interspacing i.e. microphone interspacings d, $2d$, $4d$ with $d = 0.02$ m, is positioned in the room. The coordinates of the microphones, the desired and interfering source are given in Table 2. The desired and noise signal are uncorrelated. The signal-to-noise ratio (SNR) at the first microphone equals 0 dB. To get an idea of the average performance based on directivity only, stationary desired and noise signals with the same, average long-term speech power spectral density are used. The signals can be found on [12]. Many acoustic interference signals (e.g. car noise, speech interference) have their power concentrated at the lower frequencies. Hence, their spectrum is similar to the average long-term speech spectrum. The total duration of the input signals is 10 seconds. During the first 5 seconds only noise is present, during the last 5 seconds the desired and noise signal are both present. The reverberation time T_{60} (in s) of the room is given by Eyring's formula

$$T_{60} = \frac{0.163V}{-S\ln(1-\alpha)} = \frac{0.163V}{-S\ln(\beta)}, \qquad (17)$$

with α and β the average absorption and reflection coefficient of the walls respectively, S the area (in m^2) and V the volume of the room (in m^3).
In the simulations, the recursive GSVD based algorithm is used. The exponential weighting factors λ_s and λ_n are set to 1. This gives an idea of the optimal performance for the stationary signals used. In practice, the position and the average power spectral density of the speech and noise sources vary with time and hence λ_s and λ_n should be smaller than 1. The variations with time are assumed to be slow (in the order of seconds), such that a λ_s, λ_n close to 1 can be used. It should be noted that an exponential weighting factor λ in the fullband algorithm corresponds to using a factor λ^D, with D the downsampling factor, in the subband approach.

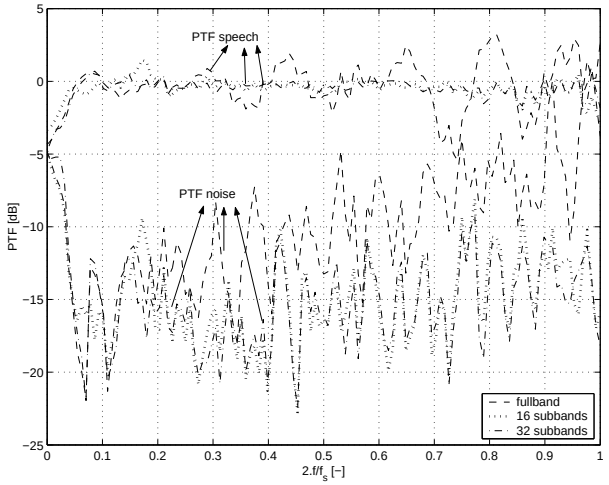


Figure 6: Power Transfer Function (PTF) of speech and noise signal for a temporal window length $T = 24$ and $T_{60} = 100$ ms, $d_f = d_g = 1$, $f_s = 16000$ Hz.

5.2 RESULTS

5.2.1 *Performance as a function of temporal window length T*

Figure 4 and figure 5 plot the improvement in SNR, ΔSNR, as well as in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$ with respect to the SNR and $\text{SNR}_{\text{intellig}}$ respectively, of the signal received at the first microphone as a function of the temporal window length T. The results are shown for 2 reverberation times T_{60}: $T_{60} \approx 100$ ms and $T_{60} \approx 300$ ms. For comparison, the performance obtained with fullband GSVD using a perfect prewhitening operation is given too. By whitening the input signals prior to GSVD filtering, the MSE is equally weighted at all frequencies.
For all T, the subband approach yields a better $\Delta\text{SNR}_{\text{intellig}}$ than the fullband approach without prewhitening and an equal or better $\Delta\text{SNR}_{\text{intellig}}$ than fullband

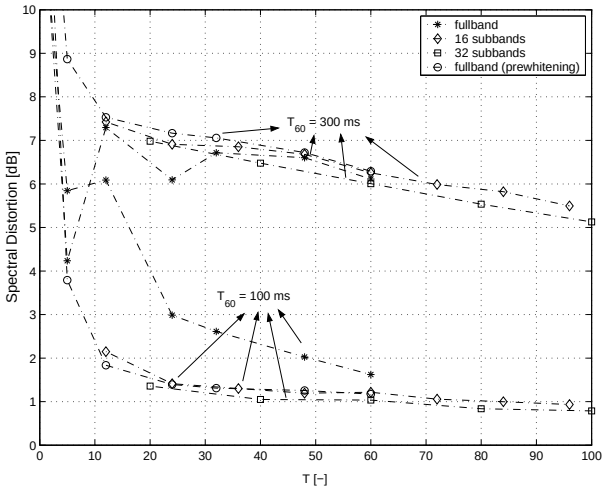


Figure 7: Spectral distortion, defined as the absolute value of the speech PTF in dB averaged over frequency, as a function of temporal window length T of fullband (with and without prewhitening) and subband approaches, $d_f = d_g = 1$.

GSVD with prewhitening. For some T a better ΔSNR is achieved too. The additional gain in $\text{SNR}_{\text{intellig}}$ offered by the subband implementation decreases if the number of degrees of freedom increases. It also decreases if the amount of reverberation grows. In the presence of larger reverberation ($T_{60} = 300$ ms) or multiple noise sources, the microphone signals become uncorrelated at high fd with f the frequency and d the microphone interspacing. The amount of uncorrelated noise reduction equals $10\log_{10}(M)$, where M is the number of microphones. Since in this case the largest amount of noise suppression is achieved at the low frequencies, the gain in $\Delta\text{SNR}_{\text{intellig}}$ is small.
Two side remarks have to be added. The absorption coefficient α in (17) (and thus the reverberation time T_{60}) is frequency dependent [13]: for some materials (e.g. carpets, curtains) the absorption coefficient α is higher at high frequencies. For real-world recordings, the improvement at low fd may be smaller due to the presence of sensor noise. At low fd, large filter weights are needed to suppress correlated noise. The presence of sensor noise limits the filter weights and hence, the performance at low frequencies.

5.2.2 *Spectral distortion*

Figure 6 depicts the Power Transfer Function (PTF) of the speech and noise signal from the first microphone to the output of the GSVD based filter obtained with the subband and fullband approaches for $T_{60} = 100$ ms and $T = 24$. A subband implementation clearly results in less distortion and larger noise reduction at the higher frequencies compared to the fullband GSVD procedure, while maintaining the performance at the low frequencies. Figure 7 plots the spectral distortion of the desired signal as a function of tem-

poral window length T. It is defined as the absolute value of the speech PTF $G_s(f)$ (in dB) averaged over frequency, i.e.

$$\frac{2}{f_s}\int_0^{\frac{f_s}{2}} |10\log_{10}(G_s(f))| \approx \frac{1}{N_f}\sum_{k=0}^{N_f-1}\left|10\log_{10}\left(G_s(k\frac{f_s}{2N_f})\right)\right| \quad (18)$$

where N_f is the number of (sample) frequencies. For $T_{60} = 100$ ms, the subband approaches and the fullband procedure with prewhitening result in a smaller distortion. For $T_{60} = 300$ ms, the distortion is larger. The GSVD based filter minimizes the MSE and is thus a compromise between noise reduction and distortion. The optimal filter attenuates the frequencies where the achievable noise reduction is low. For $T_{60} = 300$ ms, this is the case at the high frequencies. This attenuation makes a comparison between the distortion caused by the different approaches based on (18) difficult.

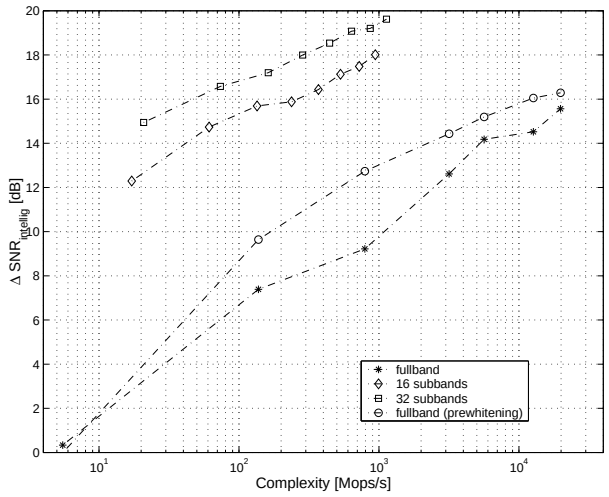


Figure 8: Improvement in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$, as a function of complexity (in Mops/s) for the subband and fullband (with and without prewhitening) approaches and $T_{60} = 100$ ms, $d_f = d_g = 1$.

5.2.3 *Improvement in $\Delta SNR_{intellig}$ as a function of complexity*

Figure 8 and figure 9 give the improvement $\Delta\text{SNR}_{\text{intellig}}$ obtained with the subband and fullband procedure as a function of complexity (expressed in Mops/s) for $T_{60} = 100$ ms and $T_{60} = 300$ ms respectively. For both reverberation times, a significant increase in $\Delta\text{SNR}_{\text{intellig}}$ is obtained: for $T_{60} = 100$ ms, an increase of 8 to 12 dB is achieved by using the subband approaches, for $T_{60} = 300$ ms, the increase is 2.5 to 5.5 dB.

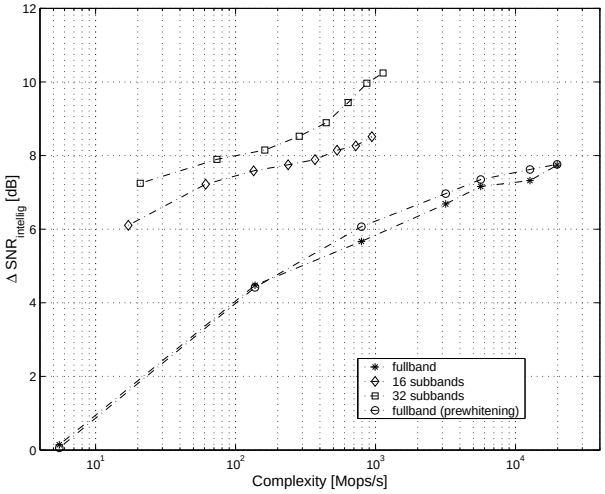


Figure 9: Improvement in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$, as a function of complexity (in Mops/s) for the subband and fullband (with and without prewhitening) approaches and $T_{60} = 300$ ms, $d_f = d_g = 1$.

6 VALIDATION OF SIMULATION RESULTS THROUGH REAL DATA EXPERIMENTS

In section 5, simulation results show that the subband GSVD approach results in a higher improvement in $\text{SNR}_{\text{intellig}}$ than the fullband algorithm, especially if the reverberation time T_{60} is small. The simulation results are validated in this section by real data experiments.

6.1 SETUP

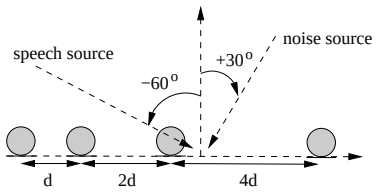


Figure 10: Position of speech and noise source in the real data experiments.

A broadside 4-microphone array with omnidirectional microphones (Knowles EM-4368) and logaritmic microphone interspacings $d, 2d, 4d$ with $d = 0.02$ m, is placed in an office room. The reverberation time T_{60} of the room is about 700 ms for a speech weighted noise signal [12]. The desired and interfering source are positioned at a distance of 1 meter from the center of the microphone array, at an angle of $-60°$ and $30°$, respectively, relative to the perpendicular of the microphone array axis (see also figure 10).The desired signal consists of three english sentences spoken by a male person. The noise signal is a stationary speech weighted noise signal with the average

long-term spectrum of the used speech material [14]. Both signals have a level of 70 dB SPL at the center of the microphone array. In the fullband GSVD algorithm, the exponential weighting factors λ_s and λ_n are set to 0.99998, in the subband algorithm λ_s and λ_n equal 0.99998^D with D the downsampling ratio. This means that approximately $\frac{1}{1-\lambda^2}$ samples at the GSVD rate, with $\lambda = \lambda_s$ and $\lambda = \lambda_n$ respectively, are used to estimate the speech + noise and noise-only second order statistics. For the chosen λ_s, λ_n, this corresponds to using two seconds of speech + noise and two seconds of noise-only data.

6.2 EXPERIMENTAL RESULTS

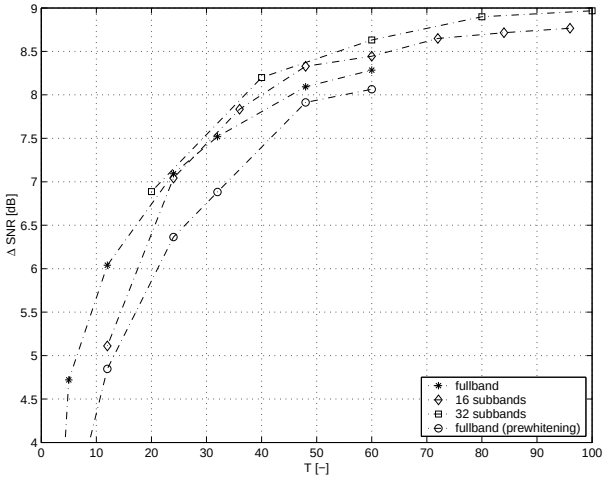


Figure 11: Improvement in SNR, ΔSNR, as a function of temporal window length T for the subband and fullband (with and without prewhitening) GSVD implementation, $d_f = d_g = 1$; experimental data.

Figure 11 and figure 12 plot the improvement in SNR, ΔSNR, as well as in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$, with respect to the SNR and $\text{SNR}_{\text{intellig}}$ respectively, of the signal received at the first microphone as a function of the temporal window length T. The performance obtained with fullband GSVD using a perfect prewhitening operation is depicted too. The experimental results confirm the simulation results: for all T, the subband approach yields a better $\Delta\text{SNR}_{\text{intellig}}$ than the fullband approach without prewhitening and an equal or better $\Delta\text{SNR}_{\text{intellig}}$ than the fullband algorithm with prewhitening, for some T a better ΔSNR is obtained too.

In figure 14, the spectral distortion is plotted as a function of the temporal window T. The subband approaches and the fullband approach with prewhitening result in a lower distortion than the fullband approach without prewhitening, especially at low T. Figure 13 depicts the PTF of the speech and noise signal from the first microphone to the output of the GSVD based filter obtained with the fullband and subband approaches for $T \approx 24$. The subband implementations lead to a bit less distortion and more noise reduction at higher frequencies. Listening to the signals reveals that the residual noise in the enhanced signals sounds more high frequent in the output of the fullband approach than in the output of the subband approach. None of the enhanced signals suffers from musical noise: since the second order statistics of speech + noise and noise-only are estimated using some seconds (e.g. 2 s) of data, the GSVD based filters do not change abruptly with time. No audible distortion is present in the enhancements.

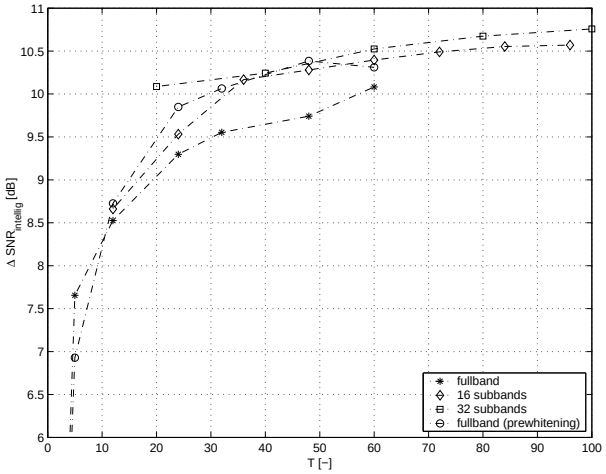


Figure 12: Improvement in intelligibility weighted SNR, $\Delta\text{SNR}_{\text{intellig}}$, as a function of temporal window length T for the subband and fullband (with and without prewhitening) GSVD implementation, $d_f = d_g = 1$; experimental data.

7 CONCLUSIONS

In speech communication applications, the presence of noise (and reverberation) in the environment may render the captured speech unintelligible. Noise reduction algorithms are thus needed to increase the intelligibility. Recently, a multi-channel noise reduction technique, based on a GSVD has been proposed. The GSVD based filter minimizes the MSE between the desired signal portion in the received signals and the sum of the filtered received microphone signals.

Many acoustic interferers have, like the average desired speech signal, their power concentrated at the low frequencies. Minimizing the fullband MSE hence pays more attention to the low frequencies than to the higher frequencies. In this paper, we propose a subband implementation of the GSVD based filter. Simulation and experimental results show that - in case of coloured signals - the subband implementation improves intelligibility and reduces distortion more than the fullband approach. In addition, a significant complexity reduction is achieved, making low cost, real-time multi-channel applications feasible.

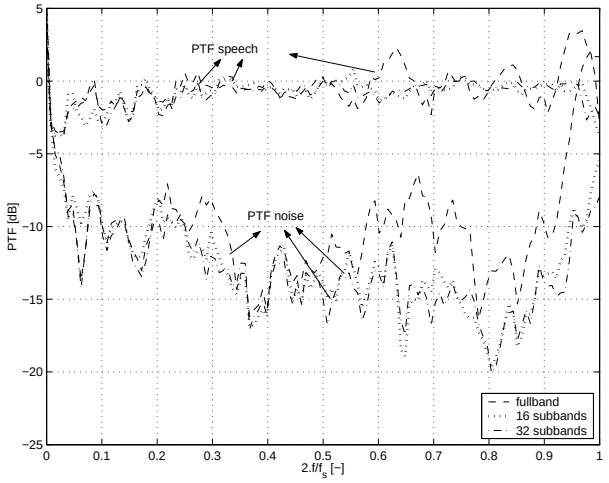


Figure 13: Power Transfer Function (PTF) of speech and noise signal for a temporal window length $T = 24$, $d_f = d_g = 1$, $f_s = 16000$Hz; experimental data.

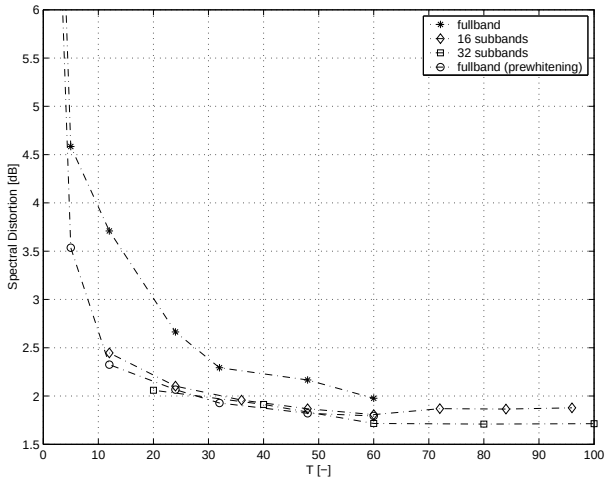


Figure 14: Spectral distortion, defined as the absolute value of the speech PTF in dB averaged over frequency, as a function of temporal window length T of fullband (with and without prewhitening) and subband approaches, $d_f = d_g = 1$; experimental data.

8 ACKNOWLEDGEMENTS

Ann Spriet is a Research Assistant supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) - Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the K. U. Leuven, in the framework of the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government, IUAP P4-02 (1997-2001) 'Modeling, Identification, Simulation and Control of Complex Systems' and FWO Research Project nr. G.0233.1 ('Signal processing and automatic patient fitting for advanced auditory prostheses').

Manuscript received on ...

REFERENCES

[1] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. In *IEEE Trans. Antennas Propag.*, Vol. 30, No.1, pages 27–34, January 1982.

[2] B. D. Van Veen and K. M. Buckley. Beamforming: A Versatile Approach to Spatial Filtering. In *IEEE ASSP Magazine*, Vol. 5, No. 2, pages 4–24, April 1988.

[3] S. Doclo and M. Moonen. *GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement*, chapter 6 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pages 111–132. Springer-Verlag, May 2001.

[4] A. Spriet, M. Moonen, and J. Wouters. Robustness analysis of GSVD based optimal Filtering and generalized Sidelobe Canceller for Hearing Aid Applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, pages 31–34, October 2001.

[5] J. E. Greenberg, P. M. Peterson, and P. M. Zurek. Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. In *Journal of the Acoustical Society of America*, Vol. 94, No. 5, pages 3009–3010, November 1993.

[6] Acoustical Society of America. ANSI S3.5-1997 American National Standard Methods for Calculation of the Speech Intelligibility Index, June 1997.

[7] M. J. Link and K. M. Buckley. Prewhitening for intelligibility gain in hearing aid arrays. In *Journal of the Acoustical Society of America*, Vol. 93, No. 4, pages 2139–2140, April 1993.

[8] K. Eneman and M. Moonen. DFT modulated Filter Bank Design for Oversampled Subband Systems. In *Signal Processing*, Vol. 81, No. 9, pages 1947–1973, September 2001.

[9] S. Weiss and R. W. Stewart. Performance Limitations of Subband Adaptive Filter. In *Proc. European Signal Processing Conference (EUSIPCO)*, Rhodos, Greece, pages 1245–1248, September 1998.

[10] M. A. Stone and C. J. Moore. Tolerable Hearing Aid Delays. I. Estimation of Limits Imposed by the Auditory Path Alone Using Simulated Hearing Losses. In *Ear and Hearing*, Vol. 20, No. 3, pages 182–191, 1999.

[11] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. In *Journal of the Acoustical Society of America*, Vol. 65, No.4, pages 943–950, April 1979.

[12] International Collegium of Reabilitative Audiology. Noise Signals ICRA Compact Disc, Ver 0.3 1997.

[13] L. L. Beranek. *Acoustics*. Acoustical Society of America, 975 Memorial Drive, Suite 804, Cambridge, MA 02138, 1993 edition, 1993.

[14] M. Nilsson, S. D. Soli, and A. Sullivan. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. In *Journal of the Acoustical Society of America*, Vol. 95, No.2, pages 1085–1099, February 1994.