IDIAP RESEARCH REPORT

# Filter Bank Design for Subband Adaptive Beamforming and Application to Speech Recognition


Kenichi Kumatani [a] John McDonough [b]
Stefan Schacht [b] Dietrich Klakow [b]
Philip N. Garner [a] Weifeng Li [a]



IDIAP–RR 08-02

February 2008



[a] IDIAP Research Institute, Martigny, Switzerland
[b] Spoken Language Systems at Saarland University in Saarbrücken, Germany



**IDIAP Research Institute** **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-02

# Filter Bank Design for Subband Adaptive Beamforming and Application to Speech Recognition

Kenichi Kumatani John McDonough Stefan Schacht Dietrich Klakow
Philip N. Garner Weifeng Li

February 2008

**Abstract.** e present a new filter bank design method for subband adaptive beamforming. Filter bank design for adaptive filtering poses many problems not encountered in more traditional applications such as subband coding of speech or music. The popular class of perfect reconstruction filter banks is not well-suited for applications involving adaptive filtering because perfect reconstruction is achieved through alias cancellation, which functions correctly only if the outputs of individual subbands are *not* subject to arbitrary magnitude scaling and phase shifts.

In this work, we design analysis and synthesis prototypes for modulated filter banks so as to minimize each aliasing term individually. We then show that the *total response error* can be driven to zero by constraining the analysis and synthesis prototypes to be *Nyquist(M)* filters.

We show that the proposed filter banks are more robust for aliasing caused by adaptive beamforming than conventional methods. Furthermore, we demonstrate the effectiveness of our design technique through a set of automatic speech recognition experiments on the multi-channel, far-field speech data from the *PASCAL Speech Separation Challenge*. In our system, speech signals are first transformed into the subband domain with the proposed filter banks, and thereafter the subband components are processed with a beamforming algorithm. Following beamforming, post-filtering and binary masking are performed to further enhance the speech by removing residual noise and undesired speech. The experimental results prove that our beamforming system with the proposed filter banks achieves the best recognition performance, a 39.6% word error rate (WER), with half the amount of computation of that of the conventional filter banks while the perfect reconstruction filter banks provided a 44.4% WER.

# 1 Introduction

There has been great interest in subband adaptive processing applications such as acoustic echo cancellation [1, 2, 3], blind signal separation [4] and beamforming [5, 6]. Subband adaptive filtering can reduce the computational complexity associated with time domain adaptive filters and improve the convergence property in estimating filter coefficients [7].

However, the filter bank design for subband adaptive filtering poses problems not seen in traditional applications such as speech coding [5][8]. de Haan showed in [8] that the perfect reconstruction (PR) filter banks were not suitable for beamforming applications because PR is achieved through aliasing cancellation [9, §5], which can reconstruct an input signal correctly only if the outputs of the individual subbands are *not* subject to arbitrary magnitude scaling and phase shifts. In [5], de Haan et al. proposed a method to design analysis and synthesis prototypes for modulated filter banks so as to minimize the weighted combination of the *response error* and *aliasing distortion*. The filter banks proposed in [5] are referred to as de Haan filter banks here.

In this work, we show that the response error defined in [5] can be driven to null by constraining the analysis and synthesis prototypes to be *Nyquist(M)* filters [9, §4.6.1]. Thereafter, the minimization of the aliasing distortions is shown to reduce to the solution of an eigenvalue problem in the case of the analysis prototype, and to the solution of a set of linear equations in the case of the synthesis prototype. We also discuss the performance limitation of the filter banks due to numerical problems, and propose an alternate solution for a special case for which we can eliminate not only the total response error but also residual aliasing distortion completely.

Those modulated filter banks are applied to a speech separation system that extracts a target speech signal. In our system, discrete-time signals are first transformed into the subband domain with our filter banks, and the subband components are then processed with the *minimum mutual information* (MMI) beamforming algorithm which estimates the *active weight vectors* so that the mutual information of both beamformer outputs is minimized [6]. Following MMI beamforming, a variant of the Wiener post-filter as well as a binary mask are applied to further reduce the residual noise.

We show the effectiveness of our filter banks through speech recognition experiments on the far-field speech data from the *PASCAL Speech Separation Challenge*. The data were recorded in a reverberant room, not artificially convoluted with measured room impulse responses. This implies that the position of the speaker's head as well as the speaking volume vary constantly.

The balance of this work is organized as follows. In Section 2 we review the definition of the uniform DFT filter bank and introduce the notation to be used throughout this work. Most importantly, we derive expressions for the total response error and residual aliasing distortion that will subsequently be minimized. In Section 3, we consider the design of suitable analysis and synthesis prototypes for the modulated filter banks discussed in Section 2. In particular, Sections 3.1 and 3.2 briefly present the prototypes design methods of [5], and then show how slight modifications of those techniques can produce prototypes with zero response error and the minimal aliasing distortion. In Section 3.3, we also present an alternate method which provides null residual aliasing distortion as well as zero total response error in a special case. In Section 4, we analyze our filter banks and compare the total response errors and aliasing distortions obtained with them to those obtained with the de Haan filter bank design. Section 5 describes beamforming techniques to which our filter banks are to be applied. In Section 6, we first describe the configuration of the automatic speech recognition system used in our experiments, and then compare our design techniques with that originally proposed in [5] as well as the popular paraunitary PR design through a set of speech recognition experiments on the multi-channel, far-field speech data from the *PASCAL Speech Separation Challenge*. Finally, in Section 7 we present our conclusions and plans for future work.

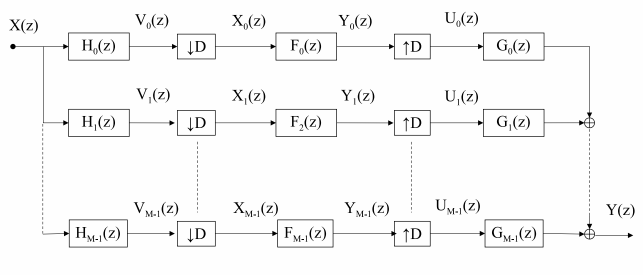


Figure 1: Schematic of a modulated subband analysis-synthesis filter bank.

# 2 Modulated Filter Banks

Figure 1 shows a schematic of a *modulated filter bank* with $M$ *subbands* and a *decimation factor* of $D$. The impulse responses $h_m[n]$ of the analysis filters are obtained by modulating a single prototype $h[n]$ according to

$$h_m[n] = h[n]\, W_M^{-mn} \leftrightarrow H_m(z) = H(zW_M^m), \tag{1}$$

where $W_M = e^{-j2\pi/M}$ denotes the $M$-th root of unity. In the popular *cosine modulated filter banks* [9, §8], the impulse responses $g_m[n]$ of the synthesis filters are time-reversed versions of the analysis filters:

$$g_m[n] = h_m[L_{\mathbf{h}} - n] \tag{2}$$

where $L_{\mathbf{h}} = mM$ is the total length of the filter bank prototype for some positive integer $m$. Provided the initial prototype $h[n]$ possesses a *paraunitary property* [9, §6.1.1], the design (2) provides *perfect reconstruction* (PR). As noted in [5][8], however, PR is achieved through alias cancellation, which functions properly only when the outputs of the individual subbands are *not* subject to arbitrary magnitude scalings and phase shifts. Hence, the PR design is not suitable for beamforming and adaptive filtering.

Following [5], we will define a separate prototype $g[n]$ for the synthesis bank, and stipulate that the individual prototypes $g_m[n]$ are related to $g[n]$ according to

$$g_m[n] = g[n]\, W_M^{-mn} \leftrightarrow G_m(z) = G(zW_M^m). \tag{3}$$

The outputs $V_m(z)$ of the analysis filters can be expressed as

$$V_m(z) = H_m(z)X(z) = H(zW_M^m)\, X(z). \tag{4}$$

Then the decimators expand the spectrums [9, §4.2] according to

$$\begin{aligned} X_m(z) &= \frac{1}{D}\sum_{d=0}^{D-1} V_m(z^{1/D}W_D^d) \\ &= \frac{1}{D}\sum_{d=0}^{D-1} H(z^{1/D}W_M^m W_D^d)X(z^{1/D}W_D^d). \end{aligned} \tag{5}$$

The last equation indicates that $X_m(z)$ consists of the sum of a stretched output of the $m$th filter bank and $D-1$ aliasing terms.

At this point, the "fixed" subband weights $F_m$ can be applied to the decimated signals to achieve the desired adaptive filtering effect:

$$Y_m(z) = F_m X_m(z). \tag{6}$$

The expanders then compress the signals $Y_m(z)$ according to

$$U_m(z) = Y_m(z^D) = \frac{1}{D} F_m \sum_{d=0}^{D-1} H(zW_M^m W_D^d) X(zW_D^d). \tag{7}$$

In the last step, the signals $U_m(z)$ are processed by the synthesis filters $G_m(z)$ in order to suppress the spectral images created by expanders, and the outputs of the synthesis filters are summed together according to

$$Y(z) = \sum_{m=0}^{M-1} U_m(z)\, G_m(z). \tag{8}$$

The final relation between the input and output signals can be expressed as

$$Y(z) = \frac{1}{D} \sum_{d=0}^{D-1} X(zW_D^d) \sum_{m=0}^{M-1} F_m\, H(zW_M^m W_D^d)\, G(zW_M^m). \tag{9}$$

Upon defining

$$A_{m,d}(z) = \frac{1}{D} F_m\, H(zW_M^m W_D^d)\, G(zW_M^m) \tag{10}$$

the output relation (9) can be written more conveniently as

$$Y(z) = \sum_{d=0}^{D-1} A_d(z)\, X(zW_D^d) \tag{11}$$

where

$$A_d(z) = \sum_{m=0}^{M-1} A_{m,d}(z). \tag{12}$$

The transfer function $A_0(z)$ produces the desired signal, while the remaining transfer functions $A_d(z)$ for $d = 1, \ldots, D-1$ give rise to the residual aliasing distortion in the output signal.

# 3 Prototype Design

Here we summarize the design methods for the analysis and synthesis prototypes proposed by de Haan *et al.* [5]. We also discuss how these methods can be modified so as to provide zero total response error while minimizing the residual aliasing distortion.

## 3.1 Analysis Prototype Design

In order to design the analysis prototype $h[n]$, de Haan *et al.* [5] define the objective function

$$\epsilon_{\mathbf{h}} = \alpha_{\mathbf{h}} + \beta_{\mathbf{h}} \tag{13}$$

where the *passband response error* is

$$\alpha_{\mathbf{h}} = \frac{1}{2\omega_p} \int_{-\omega_p}^{\omega_p} \left| H(e^{j\omega}) - H_{\mathrm{d}}(e^{j\omega}) \right|^2 d\omega, \tag{14}$$

and the *inband-aliasing distortion* is given by

$$\beta_{\mathbf{h}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{d=1}^{D-1} \left| H(e^{j\omega/D} W_D^d) \right|^2 d\omega. \tag{15}$$

In (14) the *desired filter bank response* $H_{\mathrm{d}}(e^{j\omega})$ is assumed to correspond to a pure delay of $\tau_H$ samples, such that

$$H_{\mathrm{d}}(e^{j\omega}) = e^{-j\omega\tau_H}. \tag{16}$$

Defining

$$\mathbf{h} = \left[h[0] \quad h[1] \quad \cdots h[L_{\mathbf{h}}-1]\right]^T, \tag{17}$$

$$\phi_{\mathbf{h}}(z) = \left[1 \quad z^{-1} \quad \cdots z^{-(L_{\mathbf{h}}-1)}\right]^T, \tag{18}$$

they then demonstrate that the passband response error can be expressed as

$$\alpha_{\mathbf{h}} = \mathbf{h}^T\mathbf{A}\mathbf{h} - 2\mathbf{h}^T\mathbf{b} + 1, \tag{19}$$

where

$$\mathbf{A} = \tfrac{1}{2\omega_p}\int_{-\omega_p}^{\omega_p} \phi_{\mathbf{h}}(e^{j\omega})\phi_{\mathbf{h}}^H(e^{j\omega})d\omega, \tag{20}$$

$$\mathbf{b} = \tfrac{1}{2\omega_p}\int_{-\omega_p}^{\omega_p} Re\left\{e^{j\omega\tau_H}\phi_{\mathbf{h}}(e^{j\omega})\right\}d\omega. \tag{21}$$

Based on (20–21), the components of $\mathbf{A}$ and $\mathbf{b}$ can be expressed as

$$A_{i,j} = \frac{\sin(\omega_p(j-i))}{\omega_p(j-i)}, \tag{22}$$

$$b_i = \frac{\sin(\omega_p(\tau_H - i))}{\omega_p(\tau_H - i)}. \tag{23}$$

The inband-aliasing term (15) can be expressed as

$$\beta_{\mathbf{h}} = \frac{1}{2\pi}\sum_{d=1}^{D-1}\mathbf{h}^T\left[\int_{-\pi}^{\pi}\phi_{\mathbf{h}}\left(e^{j\frac{\omega}{D}}W_D^d\right)\phi_{\mathbf{h}}^H\left(e^{j\frac{\omega}{D}}W_D^d\right)d\omega\right]\mathbf{h}. \tag{24}$$

The last equation can be rewritten as

$$\beta_{\mathbf{h}} = \mathbf{h}^T\mathbf{C}\mathbf{h}, \tag{25}$$

where

$$\mathbf{C} = \frac{1}{2\pi}\sum_{d=1}^{D-1}\int_{-\pi}^{\pi}\phi_{\mathbf{h}}\left(e^{j\frac{\omega}{D}}W_D^d\right){\phi_{\mathbf{h}}}^H\left(e^{j\frac{\omega}{D}}W_D^d\right)d\omega. \tag{26}$$

The components of $\mathbf{C}$ can then be expressed as

$$C_{i,j} = \frac{\varphi[j-i]\,\sin\left(\frac{\pi(j-i)}{D}\right)}{\pi(j-i)} \tag{27}$$

where

$$\varphi[n] = D\sum_{k=-\infty}^{\infty}\delta[n-kD] - 1.$$

Combining all terms above, de Haan *et al.* then seek to minimize the objective function

$$\begin{aligned}\epsilon_{\mathbf{h}} &= \alpha_{\mathbf{h}} + \beta_{\mathbf{h}}\\ &= \mathbf{h}^T(\mathbf{A}+\mathbf{C})\mathbf{h} - 2\mathbf{h}^T\mathbf{b} + 1.\end{aligned} \tag{28}$$

Thus, the prototype $\mathbf{h}$ proposed in [5] must satisfy

$$(\mathbf{A}+\mathbf{C})\mathbf{h} = \mathbf{b}. \tag{29}$$

**Polyphase Components**

Any given filter function $H(z)$ can be decomposed as

$$H(z) = \sum_{l=0}^{M-1} z^{-l}\, E_l(z^M), \tag{30}$$

where

$$E_l(z) \triangleq \sum_{n=-\infty}^{\infty} e_l(n)\, z^{-n} \tag{31}$$

and

$$e_l[n] \triangleq h(Mn + l), \text{ for all } 0 \leq l \leq M - 1. \tag{32}$$

Equation (30) is known as the Type 1 polyphase representation of $H(z)$ and the set $\{E_l(z)\}$ is, by definition, composed of the Type 1 polyphase components of $H(z)$; see [9, §4.3]. The Type 1 polyphase components are very useful for the efficient implementation of a modulated *analysis* filter bank. The implementation of a modulated *synthesis* bank typically relies on the Type 2 polyphase representation:

$$H(z) = \sum_{l=0}^{M-1} z^{-(M-1-l)}\, R_l(z^M), \tag{33}$$

where the set of Type 2 polyphase components $\{R_l(z)\}$ are obtained from permutation of the Type 1 polyphase components,

$$R_l(z) = E_{M-1-l}(z). \tag{34}$$

**Nyquist($M$) Filters**

Suppose that a filter function $H(z)$ has been represented in Type 1 polyphase form, and the 0-th polyphase component is constant, such that

$$H(z) = c + z^{-1}\, E_1(z^M) + \cdots + z^{-(M-1)}\, E_{M-1}(z^M). \tag{35}$$

A filter with this property is said to be a *Nyquist(M)* or *M-th band filter* [9, §4.6.1], and its impulse response clearly satisfies

$$h[Mn] = \begin{cases} c, & n = 0 \\ 0, & \text{otherwise} \end{cases} \tag{36}$$

The definition in (35) can be generalized by assuming that

$$\begin{aligned} H(z) = cz^{-m_\mathrm{d}M} &+ z^{-1}\, E_1(z^M) + \cdots \\ &+ z^{-(M-1)}\, E_{M-1}(z^M). \end{aligned} \tag{37}$$

The impulse response of $H(z)$ must then satisfy

$$h[Mn] = \begin{cases} c, & n = m_\mathrm{d} \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

In the Appendix, it is shown that if $H(z)$ satisfies (35) with $c = 1/M$, then

$$\sum_{k=0}^{M-1} H(zW^k) = Mc = 1, \tag{39}$$

where $W = e^{-j2\pi/M}$. Hence, all $M$ uniformly shifted versions of $H(e^{j\omega})$ add up to a constant. Similarly, if $H(z)$ satisfies (37), then

$$\sum_{k=0}^{M-1} H(zW^k) = z^{-m_{\mathrm{d}}M}, \tag{40}$$

in which case, in the absence of decimation, the output of the analysis filter bank would be equivalent to the input delayed by $m_{\mathrm{d}}M$ samples.

Notice that (40) represents a much stronger condition than that aimed at by the minimization of (14), in that (40) implies the response error will vanish, not just for the pass band of a single filter, but for the entire working spectrum, including the transition bands between the passbands of adjacent filters. Hence, we can replace the term $\alpha_{\mathbf{h}}$ in the optimization criterion (13) with a constraint of the form (38), then minimize the inband-aliasing distortion (24) subject to this constraint. The inband-aliasing distortion reduces to (25), the optimization of which clearly admits the trivial solution $\mathbf{h} = \mathbf{0}$. To exclude this solution, we impose the additional constraint

$$\mathbf{h}^T\mathbf{h} = 1, \tag{41}$$

which is readily achieved through the method of *undetermined Lagrange multipliers*. We posit the modified objective function

$$f(\mathbf{h}) = \mathbf{h}^T\mathbf{C}\mathbf{h} + \lambda(\mathbf{h}^T\mathbf{h} - 1) \tag{42}$$

where $\lambda$ is a *Lagrange multiplier*. Upon setting

$$\nabla f(\mathbf{h}) = \mathbf{0},$$

we find

$$\mathbf{C}\mathbf{h} + \lambda\mathbf{h} = \mathbf{0},$$

which implies

$$\mathbf{C}\mathbf{h} = -\lambda\mathbf{h}. \tag{43}$$

Hence, $\mathbf{h}$ is clearly an eigenvector of $\mathbf{C}$. Moreover, in order to ensure $\mathbf{h}$ minimizes (25), it must be that eigenvector associated with the *smallest* eigenvalue of $\mathbf{C}$. Note that, in order to ensure that $\mathbf{h}$ satisfies either (36) or (38), we must delete those rows and columns of $\mathbf{C}$ corresponding to the components of $\mathbf{h}$ that are identically zero. We then solve the eigenvalue problem (43) for the remaining components of $\mathbf{h}$, and finally reassemble the complete prototype by appropriately concatenating the zero and non-zero components. This is similar to the construction of the *eigenfilter* described in [9, §4.6.1].

## 3.2 Synthesis Prototype Design

In order to design the synthesis prototype, de Haan *et al.* [5] take as an objective function

$$\epsilon_{\mathbf{g}}(\mathbf{h}) = \gamma_{\mathbf{g}}(\mathbf{h}) + \delta_{\mathbf{g}}(\mathbf{h}) \tag{44}$$

where the *total response error* is defined as

$$\gamma_{\mathbf{g}}(\mathbf{h}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|A_0(e^{j\omega}) - e^{-j\omega\tau_T}\right|^2 d\omega, \tag{45}$$

the total analysis-synthesis filter bank delay is denoted as $\tau_T$, and the *residual aliasing distortion* is

$$\delta_{\mathbf{g}}(\mathbf{h}) = \frac{1}{2\pi}\sum_{d=1}^{D-1}\sum_{m=0}^{M-1}\int_{-\pi}^{\pi} \left|A_{m,d}(e^{j\omega})\right|^2 d\omega. \tag{46}$$

Through manipulations similar to those used in deriving the quadratic objective criterion for the analysis filter bank, it can be shown that

$$\gamma_{\mathbf{g}}(\mathbf{h}) = \mathbf{g}^T\mathbf{E}\mathbf{g} - 2\mathbf{g}^T\mathbf{f} + 1. \tag{47}$$

The components of $\mathbf{E}$ and $\mathbf{f}$ are given by

$$E_{i,j} = \frac{M^2}{D^2} \sum_{k=-\infty}^{\infty} h^*[kM - i]\, h[kM - j] \tag{48}$$

$$f_i = \frac{M}{\pi D}\, h[\tau_T - i]. \tag{49}$$

Similarly, the quadratic form for the residual aliasing distortion is

$$\delta_{\mathbf{g}}(\mathbf{h}) = \mathbf{g}^T\mathbf{P}\mathbf{g}, \tag{50}$$

where the components of $\mathbf{P}$ are given by

$$P_{i,j} = \frac{M}{D^2} \sum_{l=-\infty}^{\infty} h^*[l + j]\, h[l + i]\, \varphi[i - j],$$

$$\varphi[n] = D \sum_{k=-\infty}^{\infty} \delta[n - kD] - 1.$$

De Haan *et al.* [5] introduced a weighting factor $v$ to emphasize either the total response error (for $0 < v < 1$) or residual aliasing distortion (for $v > 1$):

$$\epsilon_{\mathbf{g}}(\mathbf{h}) = \gamma_{\mathbf{g}}(\mathbf{h}) + v\delta_{\mathbf{g}}(\mathbf{h}) \tag{51}$$

$$= \mathbf{g}^T(\mathbf{E} + v\mathbf{P})\mathbf{g} - 2\mathbf{g}^T\mathbf{f} + 1. \tag{52}$$

Hence, their synthesis prototype $\mathbf{g}$ must satisfy

$$(\mathbf{E} + v\mathbf{P})\mathbf{g} = \mathbf{f}. \tag{53}$$

### Nyquist($M$) Constraint

As with the analysis prototype, we can now impose the Nyquist($M$) constraint on the *complete analysis-synthesis prototype* $(h * g)[n]$ such that

$$(h * g)[Mn] = \begin{cases} c, & n = m_{\mathrm{d}}, \\ 0, & \text{otherwise}, \end{cases} \tag{54}$$

in which case the total response error (45) must be identically zero. Subject to this constraint, we minimize the residual aliasing distortion (52). Satisfaction of (54) clearly reduces to a set of linear constraints of the form

$$\begin{aligned} \mathbf{g}^T\mathbf{h}_{-m+1} &= 0, \\ &\vdots \\ \mathbf{g}^T\mathbf{h}_0 &= c, \\ &\vdots \\ \mathbf{g}^T\mathbf{h}_{m-1} &= 0, \end{aligned} \tag{55}$$

where $\mathbf{h}_k$ is obtained by shifting a time-reversed version of $\mathbf{h}$ by $kM$ samples and padding with zeros as needed. Equation (55) can be rewritten as

$$\mathbf{g}^T\mathbf{H} = \mathbf{c}^T, \tag{56}$$

where

$$\mathbf{H} = [\mathbf{h}_{-m+1}, \ldots, \mathbf{h}_0, \ldots, \mathbf{h}_{m-1}], \tag{57}$$
$$\mathbf{c}^T = [0, \ldots, c, \ldots, 0]. \tag{58}$$

For the constrained minimization problem at hand, we again draw upon the method of undetermined Lagrange multipliers and formulate the objective function

$$f(\mathbf{g}) = \mathbf{g}^T\mathbf{P}\mathbf{g} + (\mathbf{g}^T\mathbf{H} - \mathbf{c}^T)\lambda, \tag{59}$$

where $\lambda = [\lambda_{-m+1}, \ldots, \lambda_0, \ldots, \lambda_{m+1}]^T$. Setting

$$\nabla f(\mathbf{g}) = 2\mathbf{P}\mathbf{g} + \mathbf{H}\lambda = \mathbf{0}, \tag{60}$$

we find

$$\mathbf{g} = -\frac{1}{2}\mathbf{P}^{-1}\mathbf{H}\lambda. \tag{61}$$

The values of the multipliers $\{\lambda_k\}$ can be determined by substituting (61) into (56) and solving

$$\lambda = -2\left(\mathbf{H}^T\mathbf{P}^{-1}\mathbf{H}\right)^{-1}\mathbf{c}. \tag{62}$$

By substituting (62) into (61), we finally obtain a synthesis prototype

$$\mathbf{g} \;\; = \;\; \mathbf{P}^{-1}\mathbf{H}\left(\mathbf{H}^T\mathbf{P}^{-1}\mathbf{H}\right)^{-1}\mathbf{c}. \tag{63}$$

### 3.3 Alternate method for a special case

The optimal prototypes can be obtained by solving (43) and (63) if the matrices $\mathbf{C}$ and $\mathbf{P}$ are not singular. When those matrices are ill-conditioned, however, a different solution is required.

Theoretically speaking, the matrices $\mathbf{C}$ and $\mathbf{P}$ should not be singular because they are positive definite. From (15) and (25), the matrix $\mathbf{C}$ clearly satisfies

$$\mathbf{h}^T\mathbf{C}\mathbf{h} \geq 0. \tag{64}$$

With (46) and (50), we also find

$$\mathbf{g}^T\mathbf{P}\mathbf{g} \geq 0. \tag{65}$$

It is obvious from (64) and (65) that the matrices $\mathbf{C}$ and $\mathbf{P}$ are positive definite unless the frequency responses of the analysis and synthesis prototypes are identically zero in the stopbands. In our cases, those matrices should be positive definite and accordingly invertible. We have observed, however, that as energy in the stopbands of the analysis and synthesis prototypes approaches zero, the matrices $\mathbf{C}$ and $\mathbf{P}$ became *computationally* singular due to the limitations of floating point accuracy. This typically occurs when the decimation factor $D$ is small compared to the length of the prototype filter $L_{\mathbf{h}}$. In those cases when $\mathbf{C}$ singular, we can denote its its nullspace as $\mathbf{C}_{\text{null}}$, which consists of those column vectors $\mathbf{q} \in \mathbf{R}^n : \mathbf{C}\mathbf{q} = \mathbf{0}$. The *singular value decomposition* (SVD) [10] can be used in order obtain a basis for the nullspace of $\mathbf{C}$. Under the SVD, $\mathbf{C}$ is decomposed into

$$\mathbf{C} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T. \tag{66}$$

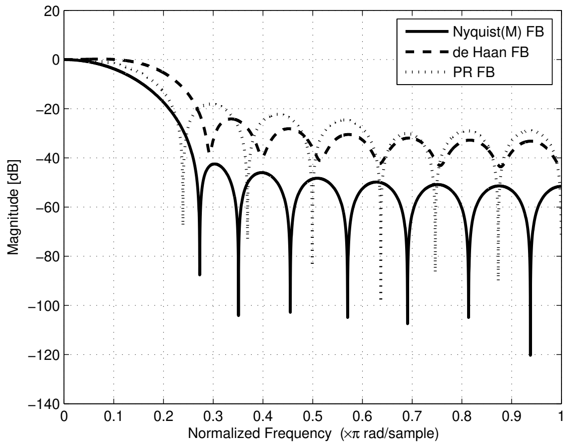


Figure 2: Frequency response of analysis filter bank prototypes with $M = 8$ subbands, decimation factor $D = 4$, and filter length $L_{\mathbf{h}} = 16$.

The bases of the null space of $\mathbf{C}$ can be obtained from the columns of $\mathbf{V}$, which correspond to singular values below a threshold. In this work, the threshold $\sigma$ is chosen such that

$$\sigma = \max(m, n) \times \max(\sigma_i) \times \epsilon$$

where $m$ and $n$ are respectively the number of rows and columns in $\mathbf{C}$, the $i$th singular value is denoted by $\sigma_i$, and $\epsilon$ floating point accuracy of the machine used for SVD computations.

Obviously, the inband-aliasing distortion can be driven to null by an analysis prototype which is represented as a linear combination of the basis vectors of the nullspace $\mathbf{h} = \mathbf{C}_{\text{null}}\ \mathbf{x}$. The free parameters $\mathbf{x}$ are determined so as to minimize the passband response error (19), the solution of which can be expressed as

$$\mathbf{h} = \mathbf{C}_{\text{null}}(\mathbf{C}_{\text{null}}^T\mathbf{A}\mathbf{C}_{\text{null}})^{-1}\mathbf{C}_{\text{null}}^T\mathbf{b}, \tag{67}$$

where rows and columns of $\mathbf{C}_{\text{null}}$, $\mathbf{A}$ and $\mathbf{b}$, corresponding to the components of $\mathbf{h}$ that are identically zero, are deleted, and $\mathbf{h}$ is reassembled so as to keep the Nyquist($M$) constraint.

For the synthesis prototype design, we can also eliminate residual aliasing distortion (50) in a similar manner. Denoting the nullspace of $\mathbf{P}$ as $\mathbf{P}_{\text{null}}$, we can express the synthesis prototype as $\mathbf{g} = \mathbf{P}_{\text{null}}\mathbf{y}$. Then by substituting into (56), we have

$$\mathbf{y} = (\mathbf{H}^T\mathbf{P}_{\text{null}})^{+}\mathbf{c} \tag{68}$$

where $(\cdot)^{+}$ indicates the peseudoinverse of $(\cdot)$. If the number of column vectors of $\mathbf{P}_{\text{null}}$ is greater than or equal to $2m-1$, we can find a synthesis prototype $\mathbf{g} = \mathbf{P}_{\text{null}}\mathbf{y}$ with zero total response error and residual aliasing distortion. We finally express the synthesis prototype with basis of the nullspace as

$$\mathbf{g} = \mathbf{P}_{\text{null}}(\mathbf{H}^T\mathbf{P}_{\text{null}})^{+}\mathbf{c}. \tag{69}$$

In practice, as the inband-aliasing distortion is very small, $\mathbf{P}$ becomes practically singular. In that case, with the method described here, we can achieve zero inband-aliasing and residual aliasing distortions.

### 3.4 Design Examples

Energy in the stopband of the filters results in aliasing. The stopband attenuation is one of the important factors to indicate how good a prototype is. Figures 2, 3 and 4 show the frequency responses of

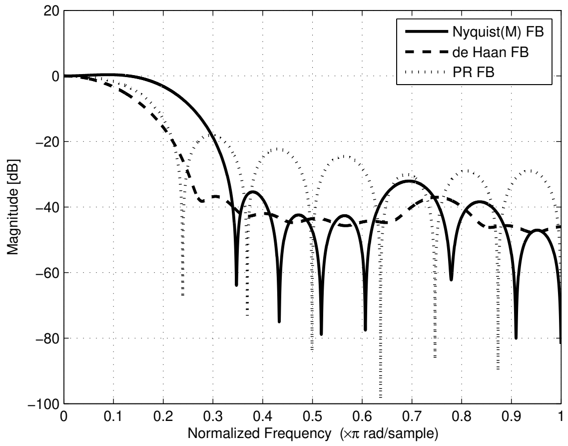


Figure 3: Frequency response of synthesis filter bank prototypes with $M = 8$ subbands , decimation factor $D = 4$, and filter length $L_{\mathbf{h}} = 16$.

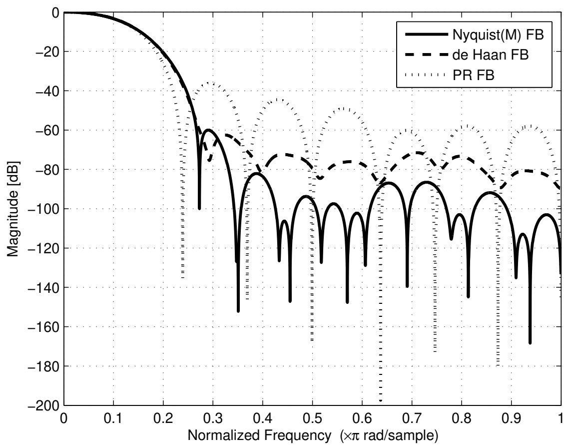


Figure 4: Frequency response of proposed composite analysis-synthesis filter bank prototypes with $M = 8$ subbands, decimation factor $D = 4$ and filter length $L_{\mathbf{h}} = 16$.

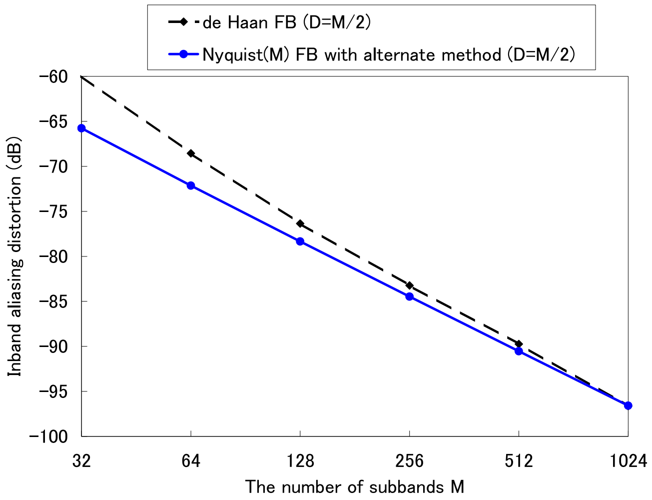


Figure 5: Inband aliasing distortion $\beta_{\mathbf{h}}$ for the number of subbands $M$. The filter length is set to $L_{\mathbf{h}} = 2M$.

the analysis, synthesis and composite analysis-synthesis prototypes respectively. Each figure presents the frequency responses of a uniform DFT filter bank using the PR prototype design (PR FB), de Haan prototype design (de Haan FB), and the proposed prototype design (Nyquist($M$) FB), where the number of subbands is $M = 8$ and the decimation factor is $D = 4$. From those figures, we can readily see that the filter banks designed by the proposed algorithm provide the highest suppression in the stopband, followed by de Haan prototype and then by the PR filter prototype. Again, in the case that arbitrary magnitude scalings and phase shifts are applied to the subband samples, the PR property is not retained. Hence, it is important to minimize the stopband energy of each filter individually.

# 4 Evaluation of Errors in Filter Prototypes

From the inband and residual aliasing distortions, we can predict the robustness of filter banks for aliasing caused by arbitrary magnitude scaling and phase shifts [8].

A relationship between the aliasing distortions and the number of subbands might be helpful for designing a beamforming system. Figures 5 and 6 show the inband and residual aliasing distortions against the number of subbands, where the decimation factor is set to $D = M/2$.

Inasmuch as decreasing the inband-aliasing distortion conduces to smaller residual aliasing distortion, it is important to minimize the inband-aliasing distortion. As shown in Figure 5, the proposed Nyquist($M$) filter prototype provides a smaller inband-aliasing distortion than the prototype designed by de Haan's algorithm, because the proposed design algorithm minimizes the inband-aliasing distortion directly while de Haan's minimize a linear combination of the passband response error and inband-aliasing distortion.

We can see from Figure 6 that the proposed algorithm can keep the residual aliasing distortion much lower than the conventional method. It is also clear from Figure 6 that the residual aliasing distortion of the Nyquist($M$) filter banks monotonically decreases with respect to the number of subbands $M$ while those of de Haan filter banks are rather invariant to it. De Haan's algorithm minimizes the linear combination of the total response error and residual aliasing distortion, equation (51). Hence, the additional term of the total response error $\gamma_{\mathbf{g}}(\mathbf{h})$ prevents the residual aliasing error $\delta_{\mathbf{g}}(\mathbf{h})$ from being suppressed. In contrast, our design technique minimizes the residual aliasing distortion only while keeping zero total response error. As a result, the residual aliasing distortion of the Nyquist($M$) filter simply decreases as $M$ increases, due mostly to the increase of the number of free parameters

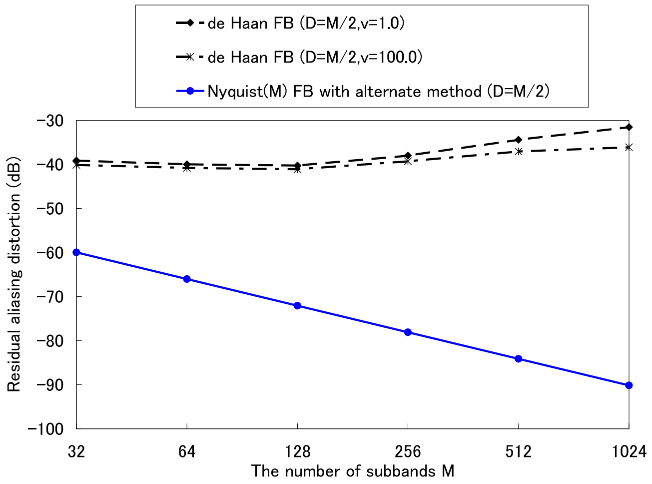


Figure 6: Residual aliasing distortion $\epsilon_{\mathbf{g}}(\mathbf{h})$ for the number of subbands $M$. The filter length is set to $L_{\mathbf{h}} = 2M$.

with respect to the number of constraints.

The aliasing errors can be also reduced by decreasing the decimation factor $D$ although it increases the computational cost associated with adaptive processing. Figure 7 presents the inband-aliasing distortions for decimation factor $D$ with de Haan's and our filter banks, where each line corresponds to a number of subbands, $M = 256$ or 512, and the filter length is set to $2M$. It is clear from Figure 7 that the proposed method suppresses the inband-aliasing distortion more than de Haan's algorithm in most cases.

In calculating the inband-aliasing distortion for the decimation factor, we observed that the matrix $\mathbf{C}$ was singular when the number of subbands and the decimation factor were set to $M = 256$ and $D \leq 32$ or $M = 512$ and $D \leq 64$. In such cases, we could find the nullspace and then use the alternate solution for the design of the analysis prototype instead of the eigen decomposition solution.

Figure 8 shows the residual aliasing distortion calculated with (50) in the same conditions as Figure 7. In Figure 8, de Haan filter banks are calculated with weighting factor $v = 100.0$. It is clear from Figure 8 that the smallest residual aliasing distortions are achieved with the proposed Nyquist($M$) filter banks. We note that when $\mathbf{C}$ was singular, $\mathbf{P}$ was also singular and the synthesis prototypes were calculated with the bases of the nullspace of the matrix $\mathbf{P}$.

Figure 9 shows the residual aliasing distortions of the Nyquist($M$) filter banks with the alternate method and without it, where the number of subbands is set to 512. As a reference, the residual aliasing distortion of the de Haan filter bank is also shown in Figure 9. In the case that the filterbanks whose decimation factor $D$ is set to less than 128 are entirely designed based on (43) and (63), the obtained solutions are instable due to the singular matrices. It is clear from Figure 9 that the residual aliasing distortion does not decrease monotonically but increases when the matrices $\mathbf{C}$ and $\mathbf{P}$ are singular. The nullspace based method can suppress the aliasing distortion even if these matrices are singular.

It could be important to know when the matrices $\mathbf{C}$ and $\mathbf{P}$ are ill-conditioned and computationally singular. We show common logarithms of *condition numbers* of those matrices in Figure 10. It is generally considered that a matrix is ill-conditioned when the condition number is too big, i.e. close to a reciprocal of floating point accuracy which is described as the threshold in Figure 10. As indicated in Figure 10, the smaller the decimation factor is set, the larger the condition number becomes. The condition numbers reach the threshold in the case of decimation factor $D \leq 64$ in Figure 10.

One might intuitively consider that the residual aliasing distortion would decrease for the decimation factor monotonically. However, Figure 8 shows that each curve of the residual aliasing distortion

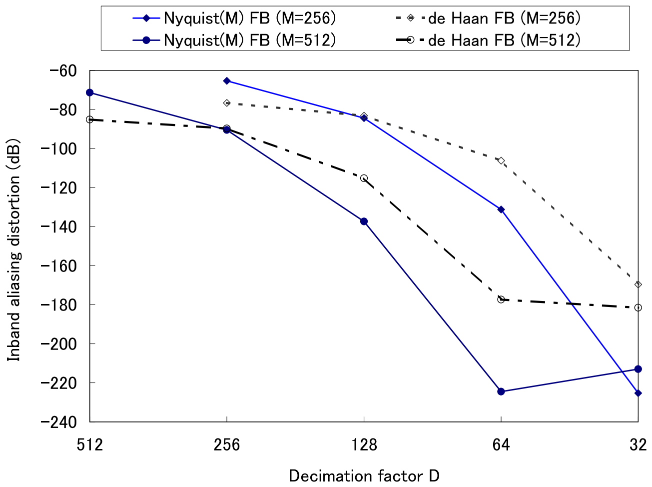


Figure 7: Inband-aliasing distortion $\beta_{\mathbf{h}}$ for decimation factor $D$. The number of subbands is $M = 256$ or $M = 512$ and the filter length is set to $L_{\mathbf{h}} = 2M$ .

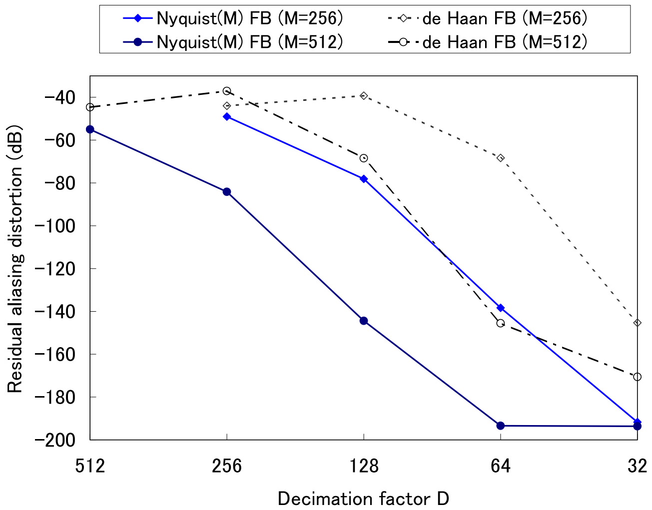


Figure 8: Residual aliasing distortion $\epsilon_{\mathbf{g}}(\mathbf{h})$ for decimation factor $D$. The number of subbands is M= 256 or M = 512 and the filter length is set to $L_{\mathbf{h}} = 2M$.

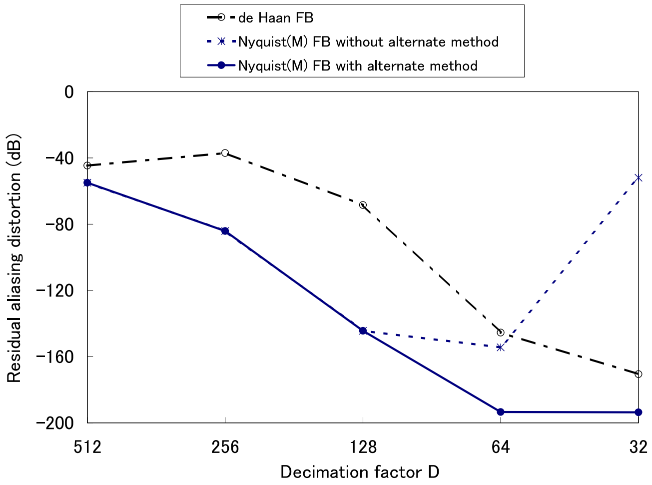


Figure 9: Comparison of the Nyquist($M$) filter banks designed with the alternate method and without it. The number of subbands is M = 512 and the filter length is set to $L_{\mathbf{h}} = 2M$.

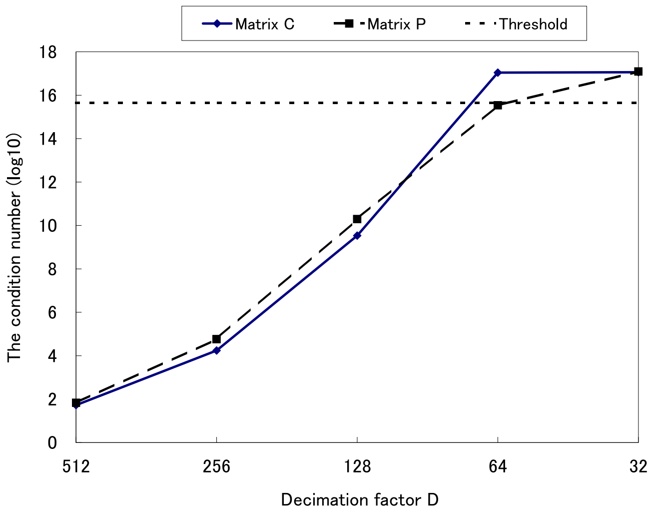


Figure 10: The common logarithm of the condition number of $\mathbf{C}$ and $\mathbf{P}$ for decimation factor $D$. The number of subbands is M = 512 and the filter length is set to $L_{\mathbf{h}} = 2M$.

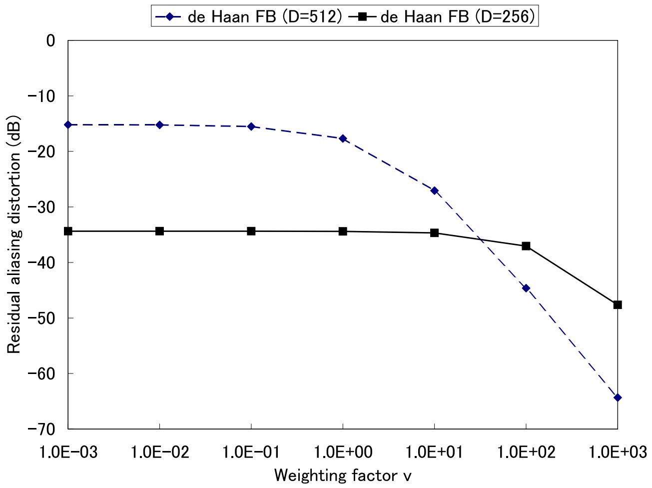


Figure 11: Residual aliasing distortion $\epsilon_{\mathbf{g}}(\mathbf{h})$ for weighting factor $v$. The number of subbands is M = 512 and the filter length is $L_{\mathbf{h}} = 2M$.

of de Haan's filter bank has a peak at $D = M/2$.

In order to look further into the reason, we calculated the residual aliasing distortions with $D$ = 256 and $D$ = 512. Figure 11 shows the residual aliasing distortions for weighting factor $v$. From Figure 11 it is seen that the residual aliasing distortion of $D = 512$ is smaller than that of $D = 256$ in the case of $v \geq 100.0$.

We also show the total response errors for weighting factor $v$ in Figure 12. It is clear from Figure 11 and Figure 12 that the residual aliasing distortion can be reduced by setting a large weighting factor $v$ at the expense of the total response error. Notice that the total response error is zero in the Nyquist($M$) filter bank.

Since the Nyquist($M$) filter banks achieve zero total response error and their residual aliasing distortion can be driven down below machine precision through a suitable selection of the decimation factor, it could be that such filter banks provide reconstruction that is "perfect" up to machine precision . In order to investigate this possibility, we calculated the mean square (MS) error $\epsilon_{\text{MS}}$ between the input $x[n]$ and output $y[n]$ of the filter bank normalized by the MS amplitude of the input, which can be expressed as

$$\epsilon_{\text{MS}} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} (x[n] - y[n])^2}{\sum_{n=0}^{N-1} x^2[n]}.$$

Figure 13 shows the MS errors of the PR, Nyquist($M$), de Haan filter banks for the decimation factor. Of course, the PR filter bank can reconstruct an exact input signal through the aliasing cancellation. Therefore, as shown in Figure 13, the PR filter bank provides the smallest MS error. The error of the PR filter bank is mainly because of the round-off error. We can also see from Figure 13 that the MS errors of the Nyquist($M$) filter banks with $D \leq 128$ are negligibly small. The total response error of de Haan's filter banks can be decreased by setting the small weighting factor $v$. However, even if $v$ is set to 0.01, as indicated in Figure 13, its MS error is the highest of the three filter banks.

Amplification of a signal would make no difference for automatic speech recognition (ASR), given that ASR front-ends all apply gain control. Therefore we also consider the normalized MS error which is invariant to such a scaling can be expressed as

$$\epsilon_{\text{Norm MS}} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} \left( x[n] - y[n] \sqrt{\frac{\sum_{n=0}^{N-1} x^2[n]}{\sum_{n=0}^{N-1} y^2[n]}} \right)^2}{\sum_{n=0}^{N-1} x^2[n]}.$$

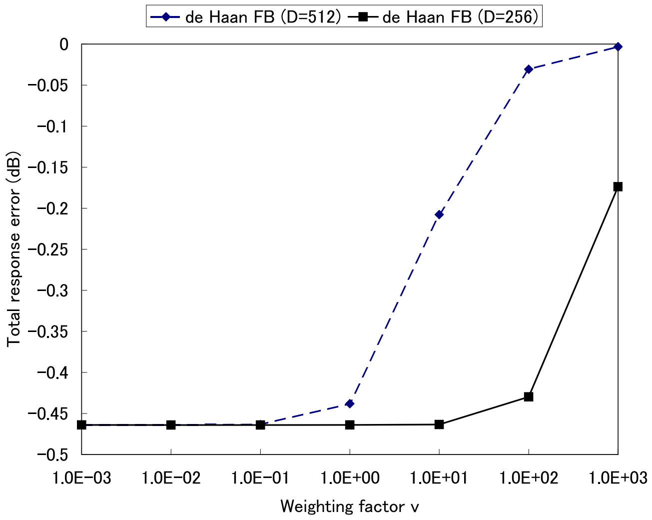


Figure 12: Total response error $\gamma_{\mathbf{g}}(\mathbf{h})$ for weighting factor $v$. The number of subbands is M $= 512$ and the filter length is $L_{\mathbf{h}} = 2M$.

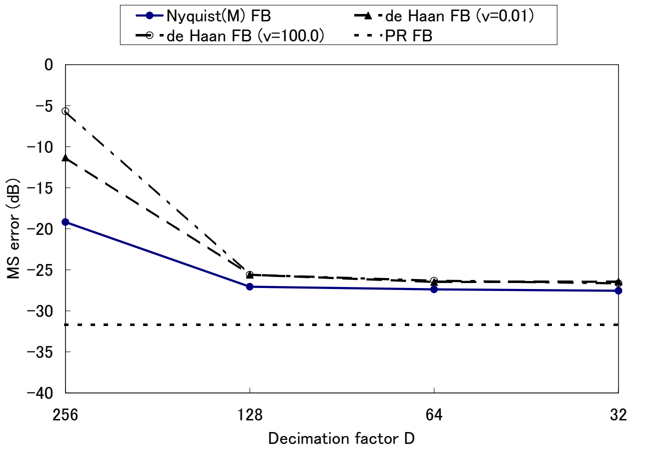


Figure 13: Mean square error (dB) for the decimation factor $D$, where $M$=512

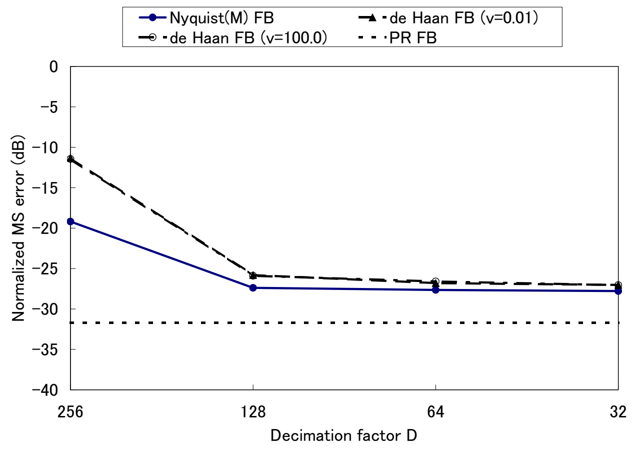


Figure 14: Normalized mean square error (dB) for the decimation factor $D$, where $M$=512

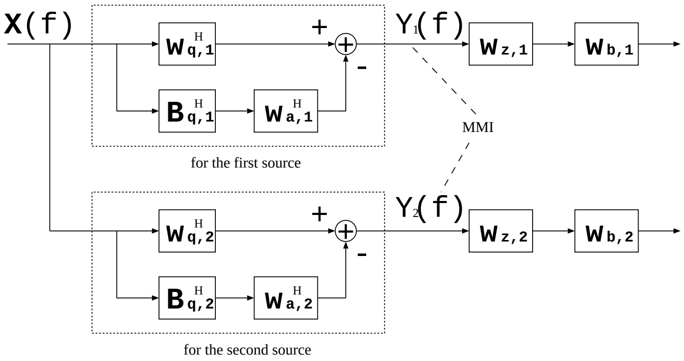


Figure 15: Schematic of generalized sidelobe cancelling (GSC) beamformers for each active source.

In this measure, the output signal is scaled so that its MS amplitude is equivalent to that of the input. Figure 14 shows the normalized MS errors of the PR, Nyquist($M$), de Haan filter banks for the decimation factor. Just as in Figure 13, the de Haan filter bank provides the worst MS error, followed by the Nyquist($M$) filter bank, then by the PR filter bank. It can also seen from Figure 14 that the weighting factor $v$ of de Haan's filter bank has no impact on the normalized MS error. This would seem to suggest that it is better to set the weighting factor $v$ to a large value for ASR.

# 5 MMI Subband Beamforming with Post-filtering

*Minimum mutual information* (MMI) beamforming can separate mixed noisy speech without the signal cancellation problems seen in conventional techniques [6]. We first transform speech signals into the subband domain with the filter banks, and then MMI beamforming is performed in the subband domain in order to extract a target signal. Consider a subband beamformer in GSC configuration [11, §6.7.3]. Assuming there are two such beamformers aimed at different sources, as shown in Fig. 15, the output of the $i$th beamformer for a given subband can be expressed as,

$$Y_i = (\mathbf{w}_{q,i} - \mathbf{B}_i \mathbf{w}_{a,i})^H \, \mathbf{X} \tag{70}$$

where $\mathbf{w}_{q,i}$ is the *quiescent weight vector* for the $i$th source, $\mathbf{B}_i$ is the *blocking matrix*, $\mathbf{w}_{a,i}$ is the *active weight vector*, and $\mathbf{X}$ is the input subband *snapshot vector*, which is common to both sources.

In keeping with the GSC formalism, $\mathbf{w}_{q,i}$ is chosen to give unity gain in the desired *look direction* [11, §6.7.3]; i.e., to satisfy a *distortionless constraint*. The blocking matrix $\mathbf{B}_i$ is chosen to be orthogonal to $\mathbf{w}_{q,i}$, such that

$$\mathbf{B}_i^H \, \mathbf{w}_{q,i} = \mathbf{0}.$$

This orthogonality implies that the distortionless constraint will be satisfied for any $\mathbf{w}_{a,i}$. While the active weight vector $\mathbf{w}_{a,i}$ is typically chosen to maximize the signal-to-noise ratio (SNR), the MMI beamforming algorithm finds the $\mathbf{w}_{a,i}$ that *minimizes* the mutual information $I(Y_1, Y_2)$. In this work, we assume that subband snapshots are Gaussian-distributed. Minimizing the mutual information criterion yields a weight vector $\mathbf{w}_{a,i}$ capable of suppressing interference that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming.

In addition to MMI beamforming, we perform *Zelinski* post-filtering [12] [13] and binary masking [14].

The Zelinski post-filter can be viewed as the extension of the single-channel *Wiener* filter to the multi-channel case. The frequency-dependent spectral weighting for the $i$th source can be expressed as

$$w_{\mathrm{z},i} = \frac{\frac{2}{M(M-1)} \left| \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \hat{\phi}_{kl} \right|}{\frac{1}{M} \sum_{k=1}^{M} \hat{\phi}_{kk}} \tag{71}$$

where $\hat{\phi}_{kk}$ is the auto-spectral density of the time-aligned input at microphone $k$ and $\hat{\phi}_{kl}$ is the cross-spectral density (CSD) at microphones $k$ and $l$. The estimation of a desired signal can be improved by averaging the CSDs.

In binary masking, we only use beamformer outputs with maximum power and set the other outputs to zero. The frequency-dependent binary masking weights $w_{\mathrm{b},i}$ can be written as:

$$\mathbf{w}_{\mathrm{b},i} = \begin{cases} 1, & \text{if } i == \text{argmax } |Y_{\mathrm{p},i}| \\ 0, & \text{otherwise} \end{cases} \tag{72}$$

where $Y_{\mathrm{p},i}$ is the frequency-dependent output of the Zelinski post-filter for the $i$th source. This masking can remove residual interference speech, but it also introduces audible artifacts into the desired speech.

# 6 Automatic Speech Recognition Experiments with Subband Beamforming

## 6.1 Configuration of speech recognition system

We performed far-field automatic speech recognition (ASR) experiments on development data from the *PASCAL Speech Separation Challenge* (SSC); see Lincoln *et al.* [15] for a description of the data collection apparatus. The data set contains recordings of five pairs of speakers where each pair of speakers reads approximately 30 sentences taken from the 5,000 word vocabulary Wall Street Journal (WSJ) task. There are a total of 43.9 minutes of speech in the development set and a total of 11,598 word tokens in the reference transcriptions. The data from two simultaneously active speakers were recorded with two circular, eight-channel microphone arrays. The diameter of each array was 20 cm, and the sampling rate of the recordings was 16 kHz. This is a challenging task for source separation algorithms given that the room is reverberant and some recordings include significant amounts of background noise. In addition, as the recorded data is real and not artificially convolved with measured room impulse responses, the position of the speaker's head as well as the speaking volume constantly vary.

Prior to beamforming, we first estimated the speaker's position with the *Orion* source tracking system [16]. In addition to the speaker's position, Orion is also capable of determining when each speaker is active. This information is very useful for speaker adaptation. We can avoid wrong speaker adaptation with interference speech when an utterance spoken by one speaker is often much longer than that spoken by the other.

Based on the average speaker position estimated for each utterance, utterance-dependent active weight vectors $\mathbf{w}_{a,i}$ were estimated for each source $i = 1, 2$. The active weights for each subband were initialized to zero for estimation with the Gaussian pdf. The snapshot covariance matrix $\Sigma_{\mathbf{X}}$ was estimated for an entire utterance. This matrix was all that was required to estimate $\{\mathbf{w}_{a,i}\}$. Thereafter iterations of the conjugate gradients algorithm were run on the entire utterance until convergence was achieved.

After beamforming, the feature extraction of our ASR system was based on cepstral features estimated with a warped *minimum variance distortionless response* [17] (MVDR) spectral envelope of model order 30. Due to the properties of the warped MVDR, neither the Mel-filterbank nor any other filter bank was needed. The warped MVDR provides an increased resolution in low–frequency regions relative to the conventional Mel-filterbank. The MVDR also models spectral peaks more accurately than spectral valleys, which leads to improved robustness in the presence of noise. Front-end analysis involved extracting 20 cepstral coefficients per frame of speech and performing global cepstral mean subtraction (CMS) with variance normalization. The final features were obtained by concatenating 15 consecutive frames of cepstral features together, then performing a *linear discriminant analysis* (LDA) to obtain a feature of length 42. The LDA transformation was followed by a second global CMS, then a global semi-tied covariance (STC) transform [18].

The far-field ASR experiments reported here were conducted with the *Millennium* automatic speech recognition system. Millenium is based on the *Enigma* weighted finite-state transducer (WFST) library, which contains implementations of all standard WFST algorithms, including weighted composition, weighted determinization, weight pushing, and minimization [19]. The *word trace decoder* in Millenium is implemented along the lines suggested by Saon *et al.* [20], and is capable of generating word lattices, which can then be optimized with WFST operations as in [21]; i.e., the raw lattice from the decoder is projected onto the output side to discard all arc information save for the word identities, and then compacted through epsilon removal, determinization, and minimization. In addition to the word trace decoder, Millenium also contains a *state trace decoder*, which maintains the full alignment of acoustic features to states during decoding and lattice generation. This state trace decoder is useful for both speaker adaptation and *hidden Markov model* (HMM) parameter estimation.

We used 30 hours of American WSJ and the 12 hours of Cambridge WSJ data in order to train a triphone acoustic model. A description of language modeling technique for dynamically composing the $HC$ and $L \circ G$ are given in [22]. Acoustic models estimated with two different HMM training schemes were used for several decoding passes: conventional maximum likelihood (ML) HMM training [23, §12], and speaker-adapted training under a ML criterion (ML-SAT) [24]. Our baseline system was fully continuous with 1,743 codebooks and a total of 67,860 Gaussian components. We performed the four decoding passes on the waveforms obtained with our speech separation systems described in previous sections. Each pass of decoding used a different acoustic model or speaker adaptation scheme. Speaker adaptation parameters were estimated using the word lattices generated during the previous pass, as in [25]. A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model and the shrunken trigram language model (LM).

2. Estimate vocal tract length normalization (VTLN) [26] parameters and constrained maximum likelihood linear regression parameters (CMLLR) [27] for each speaker, then redecode with the conventional ML acoustic model and shrunken trigram LM.

3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) [28] parameters for each speaker, then redecode with the conventional model and shrunken trigram LM.

4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SAT model and shrunken trigram LM.

Table 1: Word error rates without post-filtering for every filter bank design algorithm after every decoding passes.

| Filter bank | Pass (%WER) | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| FFT | 88.5 | 71.1 | 58.8 | 55.5 |
| PR | 87.7 | 65.2 | 54.0 | 50.7 |
| de Haan | 88.7 | 68.2 | 56.1 | 53.5 |
| Nyquist($M$) | 88.5 | 67.0 | 55.6 | 52.5 |
| CTM | 37.1 | 24.8 | 23.0 | 21.6 |

Table 2: Word error rates without post-filtering for 2 filter bank design algorithms after every decoding passes.

| Filter bank | Parameters | | Pass (%WER) | | | |
|---|---|---|---|---|---|---|
| | M | D | 1 | 2 | 3 | 4 |
| de Haan | 64 | 32 | 88.1 | 69.5 | 57.9 | 55.3 |
| | 256 | 128 | 87.3 | 69.9 | 58.2 | 54.4 |
| | 512 | 256 | 88.1 | 68.8 | 57.5 | 53.8 |
| | 512 | 128 | 87.8 | 68.9 | 56.6 | 53.7 |
| | 512 | 64 | 88.7 | 68.2 | 56.1 | 53.5 |
| Nyquist($M$) | 64 | 32 | 88.6 | 69.5 | 57.3 | 55.2 |
| | 256 | 128 | 88.0 | 70.0 | 57.1 | 54.5 |
| | 512 | 256 | 88.0 | 67.1 | 55.7 | 53.4 |
| | 512 | 128 | 88.5 | 67.0 | 55.6 | 52.5 |
| | 512 | 64 | 88.1 | 68.5 | 57.1 | 53.9 |

## 6.2 Experiments with MMI beamforming only

We first conducted speech recognition experiments on speech separated with MMI beamforming only and investigated four methods :

1. Conventional frequency domain processing based on the FFT [29],
2. Cosine modulated filter bank described by [9, §6], which yields PR under optimal conditions,
3. de Haan filter bank [5], and
4. Nyquist($M$) filter banks designed by the proposed algorithms.

Table 1 shows the word error rates (WERs) for every filter bank when we set parameters for each filter bank to obtain the best recognition performance. As a baseline, WERs for speech recorded with close-talking microphones are shown in Table 1

MMI beamforming with the PR filter banks provided the best recognition performance when post-filtering was not applied. Although it certainly scaled magnitudes and shifted phases of input subband components, we didn't observe strong aliasing noise. Hence, we were led to conclude that MMI beamforming could estimate active weight vectors while retaining aliasing cancellation. On the other hand, de Haan filter banks have a total response error which deteriorates the recognition performance. FFT analysis achieved the worst performance of all the subband processing methods.

We further investigated the recognition performance obtained with de Haan and our proposed filter banks. Table 2 depicts the WERs when different parameters for filter banks were set. In all the experiments, the filter lengths are set to twice the number of subbands, $2\times M$. It is clear from Table 2 that the proposed filter banks can provide smaller WERs than those of de Haan filter banks. These improvements are mainly because the proposed Nyquist($M$) filter banks can have zero total response error. From Table 2, one can also see that as the number of subbands $M$ increases, the WER decreases. the MMI beamforming algorithm can strengthen a target wave by using its echoes which are caused by a reflection on a hard surface such as a table [6]. Thus, the larger number of subbands generally leads to the better performance of speech enhancement of the MMI beamformer. In order to enhance this advantageous effect, we need to make the length of the analysis filter enough long to include such reflected waves in the analysis window. This can be done by increasing the number of subbands.

Contrary to our expectations, Table 2 shows that the WER of the Nyquist($M$) filter bank doesn't monotonically decrease with the decimation factor although the residual aliasing distortion does. We suppose that it is because of the numerical instability discussed in Section 4.

### 6.3 Experiments with MMI beamforming and post-filtering

Finally we did speech recognition experiments on speech enhanced with Zelinski post-filtering and binary masking after MMI beamforming. Table 3 shows WERs in those experiments. In this case, the PR property was not kept because of the rapid change of filter weights. We heard the aliasing distortions when the PR filter banks were used. In contrast, de Haan and the proposed filter banks could suppress such aliasing noise because those filter banks are designed so as to minimize aliasing terms individually.

In Table 3, unlike the trend seen in Table 2, the larger number of subbands, which leads a higher frequency resolution, doesn't necessarily provide the better recognition performance. We suppose that is due to inaccurate estimation of noise spectrums in Zelinski-postfiltering. In the case of a high frequency resolution, many many filter coefficients must be estimated for Zelinski-postfiltering, which could lead to robustness problems.

One could find correlation between the WER and the residual aliasing distortion by paying attention to the relationship between Table 3 and Figure 8. Generally, as the residual aliasing distortion is reduced, the WER becomes smaller. However, this is not always true because there are many other factors impacting recognition performance. For example, although the residual aliasing distortion of the Nyquist($M$) filter bank vanishes with an increase in the number of subbands, increasing the number of subbands can lead to robustness problems in estimating the postfiler coefficients; hence, the WER doesn't monotonically decrease for an increasing number of subbands.

Table 3 also shows that the systems with de Haan and Nyquist($M$) filter banks can reduce the absolute WER by about 5.0 % compared to those with the PR filter banks. This suggests that the PR filter bank is less suitable for adaptive processing. It is also clear from Table 3 that the proposed method achieved a bigger WER reduction than de Haan's algorithm. In particular, the improvement of the recognition performance is significant with $M = 256$. The proposed filter banks achieved the best recognition performance, WER 39.6 % with the number of subbands $M = 512$ and decimation factor $D = 128$. On the other hand, de Haan filter banks provided the same number with $M = 512$ and $D = 64$. Therefore, our method can be thought of as halving the computational cost of that of de Haan.

## 7 Conclusions

In this work, we have proposed a new design method for filter banks which is suitable for adaptive processing. We have thoroughly investigated the properties of the proposed filter banks and compared them with the filter banks proposed by de Haan through speech recognition experiments on real data from the PASCAL Speech Separation Challenge. We also proved that minimization of the residual aliasing errors is more important than the perfect reconstruction property in adaptive processing.

Table 3: Word error rates with post-filtering for every filter bank design algorithm after every decoding passes.

| Filter bank | Parameters | | Pass (%WER) | | | |
|---|---|---|---|---|---|---|
| | M | D | 1 | 2 | 3 | 4 |
| PR | 64 | - | 83.7 | 61.5 | 47.5 | 44.7 |
| | 512 | - | 84.6 | 60.5 | 47.6 | 44.4 |
| de Haan | 64 | 32 | 82.4 | 59.2 | 46.2 | 43.3 |
| | 256 | 128 | 82.0 | 60.5 | 44.7 | 42.0 |
| | 512 | 256 | 83.9 | 59.1 | 43.2 | 41.3 |
| | 512 | 128 | 81.6 | 58.9 | 43.2 | 40.3 |
| | 512 | 64 | 82.7 | 57.7 | 42.7 | 39.6 |
| Nyquist($M$) | 64 | 32 | 80.7 | 57.0 | 44.3 | 42.0 |
| | 256 | 128 | 81.0 | 56.2 | 41.8 | 39.8 |
| | 512 | 256 | 84.1 | 58.6 | 43.4 | 40.6 |
| | 512 | 128 | 81.8 | 54.9 | 42.2 | 39.6 |
| | 512 | 64 | 81.4 | 56.5 | 42.6 | 40.3 |

Furthermore, we analyzed how the parameters of filter banks influence the recognition performance through MMI beamforming and post-filtering.

The proposed method achieved the smallest WER (39.6 %) with half as much computational cost as de Haan filter banks while the PR filter provided a 44.4 % WER.

In the future, we plan to apply the proposed Nyquist($M$) filter banks to different adaptive beamforming algorithms. In particular, we will use the proposed filter banks in MMI beamforming with new super-Gaussian assumptions. We also plan to develop a new subband beamforming algorithm with the proposed filter banks. In future work, the filter banks described here will be a fundamental technique in adaptive processing and beamforming.

# A Proof of Equations (39–40)

Repeating (35) we have

$$H(z) = c + z^{-1}\,E_1(z^M) + \cdots + z^{-(M-1)}\,E_{M-1}(z^M) \tag{73}$$

so that

$$\begin{aligned}
\textstyle\sum_{k=0}^{M-1} \quad & H(zW_M^k) \\
= \quad & \sum_{k=0}^{M-1} [c + (zW_M^k)^{-1} E_1(z^M W_M^{km}) + \cdots \\
& +(zW_M^k)^{-(M-1)} E_1(z^M W_M^{km})] \\
= \quad & \sum_{k=0}^{M-1} [c + W_M^{-k} \cdot z^{-1} E_1(z^M) + \cdots \\
& +W_M^{-(M-1)k} \cdot z^{-(M-1)} E_{M-1}(z^M W_M^{km})] \\
= \quad & c \sum_{k=0}^{M-1} 1 + z^{-1} E_1(z^M) \sum_{k=0}^{M-1} W_M^{-k} + \cdots \\
& +z^{-(M-1)} E_{M-1}(z^M) \sum_{k=0}^{M-1} W_M^{-k(M-1)}
\end{aligned} \tag{74}$$

Now note that

$$\sum_{k=0}^{M-1} x^m = \frac{1-x^M}{1-x}$$

Hence,

$$\sum_{k=0}^{M-1} W_M^{-km} = \frac{1-W_M^{-mM}}{1-W_M}$$

and as $W_M^{-mM} = 1$ for all $m = 1, 2, \ldots$, it is clear that all terms in (74) save for the first vanish. Therefore, for $H(z)$ in (73) with $c = 1/M$,

$$\sum_{k=0}^{M-1} H(zW_M^k) = 1$$

Similarly, if the impulse response associated with $H(z)$ is delayed by $m_d M$ samples to obtain a causal system, then

$$H(z) = cz^{-m_d M} + z^{-1}\, E_1(z^M) + \cdots + z^{-(M-1)}\, E_{M-1}(z^M)$$

and it is readily verified

$$\sum_{k=0}^{M-1} H(zW_M^k) = z^{-m_d M}$$

which implies the composite analysis-synthesis filter bank produces a simple delay of the input in the absence of aliasing

# References

[1] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling : analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, pp. 1862–1875, 1992.

[2] P.A. Naylor, O. Tanrikulu, and A.G. Constantinides, "Subband adaptive filtering for acoustic echo control using allpasspolyphase IIR filterbanks," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 143–155, 1998.

[3] Steven L. Gay (Editor) and Jacob Benesty, *Acoustic Signal Processing for Telecommunication*, Springer, 2000.

[4] K. Kokkinakis and P. C. Loizou, "Subband-based blind signal processing for source separation in convolutive mixtures of speech," in *Proc. ICASSP*, 2007, pp. 917–920.

[5] Jan Mark de Haan, Nedelko Grbic, Ingvar Claesson, and Sven Erik Nordholm, "Filter bank design for subband adaptive microphone arrays," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 1, pp. 14–23, Jan. 2003.

[6] Kenichi Kumatani, Tobias Gehrig, Uwe Mayer, Emilian Stoimenov, John McDonough, and Matthias Wölfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2527–2541, November 2007.

[7] John J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, pp. 14–37, 1992.

[8] Jan Mark de Haan, *Filter Bank Design for Subband Adaptive Filtering*, Ph.D. thesis, Karlskrona. Blekinge Institute of Technology, 2001.

[9] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, 1993.

[10] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling, *Numerical Recipes in C*, Cambridge University Press, 1992.

[11] H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.

[12] Claude Marro, Yannick Mahieux, and Klaus Uwe Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240–259, 1998.

[13] Iain A. McCowan and Hervé Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 709–716, 2003.

[14] Iain A. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio-visual sensor array," in *in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[15] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus ( mc-wsj-av): Specification and initial experiments," in *Proc. ASRU*, 2005, pp. 357–362.

[16] Tobias Gehrig, Ulrich Klee, John McDonough, Shajith Ikbal, Matthias Wölfel, and Christian Fügen, "Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters," in *in Proc. Interspeech*, 2006, pp. 2594–2597.

[17] Matthias Wölfel and John McDonough, "Minimum variance distortionless response spectral estimation: Review and refinements," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[18] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[19] M. Mohri and M. Riley, "Network optimizations for large vocabulary speech recognition," *Speech Comm.*, vol. 28, no. 1, pp. 1–12, 1999.

[20] G. Saon, D. Povey, and G. Zweig, "Anatomy of an extremely fast LVCSR decoder," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 549–552.

[21] A. Ljolje, F. Pereira, and M. Riley, "Efficient general lattice generation and rescoring," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 1251–1254.

[22] John McDonough, Emilian Stoimenov, and Dietrich Klakow, "An algorithm for fast composition of weighted finite-state transducers," in *Proc. ASRU*, 2007.

[23] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing, New York, 1993.

[24] T. Anastasakos, J. McDonough, R. Schwarz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[25] L. Uebel and P. Woodland, "Improvements in linear transform based speaker adaptation," in *Proc. ICASSP*, 2001.

[26] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalizatio," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 6, pp. 415–426, 2002.

[27] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, 1998.

[28] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, April 1995.

[29] Futoshi Asano, Shiro Ikeda, Michiaki Ogawa, Hideki Aso, and Nobuhiko Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 3, pp. 204–215, May 2003.