SUBBAND NONSTATIONARY NOISE REDUCTION BASED ON MULTICHANNEL SPATIAL PREDICTION UNDER REVERBERANT ENVIRONMENTS

Masahito Togami, Yohei Kawaguchi, and Yasunari Obuchi

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

We propose a novel non-stationary and convolutive noise reduction method under reverberant environments. Unlike many multichannel noise reduction methods, the proposed method does not need pre knowledge of impulse response or direction of arrival (DOA) of the target source. The proposed method is composed of two processes. On the noise reduction process, the noise component is reduced without the impulse response of the target source. The target source component in the output signal is distorted, but the distortion is removed by the distortion-restoration process. Importantly, possibility of complete noise reduction with no distortion based on the proposed framework is assured by MINT theory. Experimental results under the reverberant environment ($RT_{60} \approx 300$ ms) show that the proposed method can reduce more noise than the conventional method and the distortion of the target source is not so big.

Index Terms— multichannel noise reduction, spatial prediction, subband, reverberation

1. INTRODUCTION

For conference recording systems or video conferencing systems, noise reduction technique is required. Previously, many researches have been done on noise reduction technique with multiple microphones (microphone array) [1][2]. These methods use approximated impulse response by direction of arrival (DOA) of the target source. When the direct sound from the sound source is sufficiently bigger than reverberation, conventional methods can reduce noises effectively. However, when reverberation is dominant, noise reduction performance is greatly degraded. Under the reverberant environments, the difference between modeled impulse response which is approximated by DOA of the target source and the actual impulse response is so big, and part of the target source component is reduced together with noise source component (signal cancellation problem). Commonly, video conferencing systems or recording systems are used at reverberant environments such as office rooms. Ideally, when the impulse response between the target source and a microphone array is given, noise components and reverberation can be completely removed [3]. However, in the general case, we cannot obtain the impulse response. Chen, et al. [4] proposed minimum distortion beamformer based on spatial prediction (SP-MDBF) for reverberant environments, which don't need the impulse response of the target source. Instead of estimation of the original target source component, SP-MDBF estimates the target source component at one microphone position. SP-MDBF requires only spatial prediction coefficients between the target source component at one microphone and the target source component at the other microphone. However, to obtain spatial prediction coefficients, second order statistics (SOS) of the target source is needed. Chen, et al. [4] proposed that SOS of the target source is approximated by subtraction SOS of the noise sources from SOS of the mixing signal. This approximation is correct only when SOS of the noise sources are stationary. When noise sources are non-stationary, such as human speech, Due to misestimation of SOS, the output signal is greatly distorted.

In this paper, we propose a novel noise reduction method which can reduce noise effectively at reverberant environments. The proposed method does not require the assumption that SOS of the noise sources are stationary and does not require the impulse response between the target source and a microphone array. The proposed method is composed of two processes. On the noise reduction process, the noise component at each microphone position is reduced without any constraint to the impulse response of the target source. Theoretically, noise component can be completely removed by multichannel spatial prediction filter which is an extension of the spatial prediction of SP-MDBF. However, the output signal of the noise reduction process is distorted because there is no constraint to the target signal. The distortion is removed by the distortion-restoration process from multichannel distorted target source components. Theoretically, it is assured by MINT theory [5] that we can obtain completely noiseless signal with no distortion of the target source by the proposed framework. To reduce computational cost, the proposed method is performed at subband domain, the length of the impulse response at each subband can be assumed to be shorten by downsampling and computational cost can be reduced at the realistic level. Experimental results under the reverberant environment ($RT_{60} \approx 300$ ms) show that the proposed method can reduce more noise than SP-MDBF and the distortion of the target source is not so big.

2. PROBLEM STATEMENT

2.1. Input signal model

M is defined as the number of microphones. N is described as the sum of the number of the target signals N_s and the number of the noise signals N_n. Received sound signal at the m-th microphone element is described as $x_m(t)$. t is the sampling number of an A/D converter. The original source signal of the i-th target signal is described as $s_i(t)$, and that of the i-th noise signal is described as $n_i(t)$. $h_{i,m}$ is the impulse response of the i-th target signal between the i-th source position and the m-th microphone. $g_{i,m}$ is the impulse response of the i-th noise signal. $x_m(t)$ is defined as $x_m(t) = \sum_{i=0}^{N_s-1} (h_{i,m} * s_i(t)) + \sum_{i=0}^{N_n-1} (g_{i,m} * n_i(t))$, $*$ is the operator of convolution. In this paper, the noise reduction problem is defined as extraction of the target source component $\sum_{i=0}^{N_s-1} (h_{i,c} * s_i(t))$ at the c-th microphone from M noisy input signals $[x_1(t), \ldots, x_M(t)]$ by using time period in which there is only noise sources (only-noise-period). c is defined as the *target mi-*

crophone index. For easily description, the target source component is replaced to $y_m(t) = \sum_{i=0}^{N_s-1}(h_{i,m} * s_i(t))$, and the noise source component is replaced to $v_m(t) = \sum_{i=0}^{N_n-1}(g_{i,m} * n_i(t))$. Therefore, $x_m(t) = y_m(t) + v_m(t)$. Furthermore, the vector $\boldsymbol{x}_m(t)$ is defined as $[x_m(t) \dots x_m(t-L)]^T$, $\boldsymbol{y}_m(t)$ is $[y_m(t) \dots y_m(t-L)]^T$, and $\boldsymbol{v}_m(t)$ is $[v_m(t) \dots v_m(t-L)]^T$.

2.2. Conversion into the subband domain

Under reverberant environments, even if the sampling frequency is low such as 8 kHz, the length of the impulse response exceeds a thousand tap at time domain. To reduce computational cost, subband processing framework by oversampled DFT filter bank [6] is utilized and noise reduction is performed at subband domain. To avoid aliasing problem at each subband, The downsampling rate R is set to be smaller than the number of subbands K. In this paper, K is set to be 64, and R is set to be 56 for 8 kHz sampling. $x_m(k,t)$ is defined as the k-th subband signal of the m-th microphone input signal.

3. SPATIAL PREDICTION BETWEEN A MICROPHONE PAIR

Chen, et al. [4] proposed minimum distortion beamformer using the spatial prediction between a microphone pair (SP-MDBF). The target source component in the i-th microphone is predicted from the target source component in the j-th microphone as follows:

$$\hat{\boldsymbol{y}}_i(k,t) = \boldsymbol{a}_{i \leftarrow j}(k)\boldsymbol{y}_j(k,t), \tag{1}$$

where $\boldsymbol{a}_{i \leftarrow j}(k)$ is the spatial prediction coefficient. $\boldsymbol{a}_{i \leftarrow j}(k)$ is easily calculated by second order statistics (SOS) of the target source component as $\boldsymbol{a}_{i \leftarrow j}(k) = \boldsymbol{R}_{i,j}(k)\boldsymbol{R}_j(k)^{-1}$, $\boldsymbol{R}_j(k) = E[\boldsymbol{y}_j(k,t)\boldsymbol{y}_j(k,t)^*]$, $\boldsymbol{R}_{i,j}(k) = E[\boldsymbol{y}_i(k,t)\boldsymbol{y}_j(k,t)^*]$, $*$ is the operator of conjugate transpose, and E is the operator which calculated expected value. Second order statistics of the target source component can be approximated as ensemble average of corresponding statistics calculated in time period when there is only target source (only-target-period). Therefore, spatial prediction coefficient can be obtained without the impulse response. However, because only-target-period cannot be obtained, in SP-MDBF, $\boldsymbol{R}_j(k)$ is approximated as $\boldsymbol{R}_{in,j}(k) - \boldsymbol{R}_{noise,j}(k)$, and $\boldsymbol{R}_{i,j}(k)$ is approximated as $\boldsymbol{R}_{in,i,j}(k) - \boldsymbol{R}_{noise,i,j}(k)$. $\boldsymbol{R}_{noise,j}(k)$ and $\boldsymbol{R}_{noise,i,j}(k)$ are SOS of the noise source component, and $\boldsymbol{R}_{in,j}(k)$ and $\boldsymbol{R}_{in,i,j}(k)$ are SOS of the noisy input signal. The above approximation is correct only when SOS of the noise source is stationary. Therefore, when the noise source is non-stationary such as human speech, by SP-MDBF framework, the spatial prediction of the target source component cannot be correctly estimated.

3.1. Limitation of the spatial prediction filter

Theoretically, the ideal value of the spatial prediction filter $\boldsymbol{a}_{i \leftarrow j}(k)$ is described as $h_i(k,z)h_j(k,z)^{-1}$ by the z-transformation. However, $h_j(k,z)$ is usually a non minimum phase filter, and the inverse filter cannot be obtained. Therefore, theoretically, the prediction error of $\boldsymbol{a}_{i \leftarrow j}(k)$ cannot be zero.

4. PROPOSED METHOD

The block diagram of the proposed method is shown in Fig. 1. To reduce noise without the impulse response of the target source and only-target-period, in noise reduction process, noise reduction filter is adapted with no pre knowledge of the target source. Motivated by spatial prediction technique of SP-MDBF, noise reduction is performed by spatial prediction. Predicted noise component of the m-th microphone is subtracted from the m-th noisy signal. However, the prediction error of the spatial prediction filter cannot be zero. By MINT theory [5], it is assured that even if the impulse response is a non minimum phase filter, the inverse filter can be obtained by using multichannel signals which have no common zero. Therefore, the proposed method utilizes multichannel spatial prediction to reduce prediction error of spatial prediction filter. Theoretically, prediction error can be zero by multichannel spatial prediction. However, the target signal in output signal of noise reduction process is distorted. In the proposed distortion-restoration process, distortion of the target signal in the output signal of noise reduction process is restored by multichannel restoration filter. The restoration filter is updated so as to approximate the filtered signal to the microphone input signal. When the impulse response between the original target source signal and the target signal in output signal of noise reduction process at each microphone have no common zero, the inverse filter can be obtained by using multichannel signals. Therefore, by the proposed framework, theoretically, complete noise reduction with no distortion is possible.

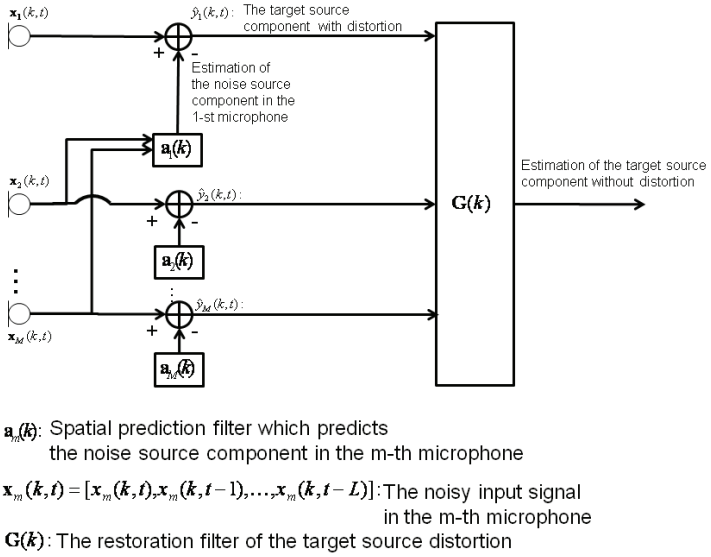


Fig. 1. Block diagram of the proposed method

4.1. Multichannel spatial prediction

The noise source component in the m-th microphone v_m is reduced by multichannel spatial prediction filter $\boldsymbol{a}_m$ as follows:

$$\epsilon_m(k,t) = v_m(k,t) - \boldsymbol{a}_m \boldsymbol{v}_m^e(k,t), \tag{2}$$

where $\boldsymbol{v}_m^e(k,t) = [\boldsymbol{v}_1(k,t)^T, \dots, \boldsymbol{v}_{m-1}(k,t)^T, \boldsymbol{v}_{m+1}(k,t)^T, \dots]^T$, and $\epsilon_m(k,t)$ is the residual noise. Theoretically, the ideal value of $\boldsymbol{a}_m$ is interpreted as multiplication of MINT inverse filter of all noise signals except for the m-th microphone signal and the impulse response of the noise source component at the m-th microphone signal. By MINT theorem [5], if the filter length of

$\boldsymbol{a}_m$ for each microphone L is larger than $\frac{N_n(L_g-1)}{M-2}$, and the impulse response of the noise source component at each microphone doesn't have common zero, MINT inverse filter exists, and the prediction error of $\boldsymbol{a}_m$ can be completely zero. $\boldsymbol{a}_m$ is obtained by second order statistics at only-noise-period as $\boldsymbol{R}_{cor,v}(m)\boldsymbol{R}_{cov,v}(m)^{-1}$, $\boldsymbol{R}_{cor,v}(m) = E[\boldsymbol{v}_m(k,t)\boldsymbol{v}_m^e(k,t)^*]$, $\boldsymbol{R}_{cov,v}(m) = E[\boldsymbol{v}_m^e(k,t)\boldsymbol{v}_m^e(k,t)^*]$.

Output signal of the noise reduction process for the m-th noisy input signal is described as follows:

$$\hat{y}_m(k,t) = x_m(k,t) - \boldsymbol{a}_m \boldsymbol{x}_m^e(k,t). \quad (3)$$

$\hat{y}_m(k,t)$ is noiseless, but is distorted because there is no constraint to the target signal in noise reduction process. The distortion is removed by the following distortion-restoration process.

4.2. Distortion-restoration process

After noise reduction process, M channel distorted target source components $\hat{y}_m(k,t)$ (from $m = 1$ to M) are obtained. $\hat{y}_m(k,t)$ is regarded as $q_m(k) * s(k,t)$. $q_m(k)$ is composed of the impulse response of the target source at the m-th microphone and the impulse response of the noise reduction filter. The ideal value of the restoration filter is multiplication of the inverse filter of $q_m(k)$ and the impulse response of the target source at the target microphone. If $q_m(k)$ has no common zero and restoration filter length in each microphone L_r is larger than $\frac{N_s(L_g+L-1)}{M-1}$, according to MINT theory, MINT inverse filter of $q_m(k)$ can be obtained. Therefore, complete noise reduction with no distortion of the target source is possible by the proposed framework. However, actually, there is the residual noise in the output signal of noise reduction process. Under the assumption that there is the residual noise, the restoration filter $\boldsymbol{G}(k)$ which minimizes the cost function of $f(\boldsymbol{G}(k))$ is used.

$$f(\boldsymbol{G}(k)) = |x_c(k,t) - \boldsymbol{G}(k)\boldsymbol{y}(k,t)|^2 + \mu|\boldsymbol{G}(k)\boldsymbol{n}(k,t)|^2, \quad (4)$$

where $\boldsymbol{y}(k,t) = [\boldsymbol{y}_1(k,t)^T, \ldots, \boldsymbol{y}_M(k,t)^T]^T$, $\boldsymbol{n}(k,t)$ is the residual noise component measured at only-noise-period and is defined as the same form of $\boldsymbol{y}(k,t)$, $\boldsymbol{y}_m(k,t) = [y_m(k,t), \ldots, y_m(k,t-L_r+1)]^T$, and ML_r is the length of the restoration filter. When $\boldsymbol{y}(k,t)$ has no residual noise, the multichannel filter $\boldsymbol{G}(k)$ which minimizes the first term can completely reduce distortion of the target source. When noise signal is remained in $\boldsymbol{\epsilon}(k,t)$, the multichannel filter $\boldsymbol{G}(k)$ which minimizes the first term is greatly large-valued. The second term protects growth of the filter $\boldsymbol{G}(k)$ [7].

Remained distortion after $\boldsymbol{G}(k)$ is restored by the additional single channel FIR filter which is adapted so as to approximate the filtered signal to the microphone input signal.

5. EXPERIMENT

The proposed method was evaluated by the experiment under the reverberant environment ($RT_{60} \approx 300$ ms). The sound for evaluation was made by the impulse responses which were recorded at the above environment. The sampling rate was 8 kHz. The target source signal was human speech. The number of the microphones in the microphone array was 12. The length of the microphone array was 70 cm. The target microphone was the seven-th microphone. The number of the target source N_s and the number of the noise source N_n were 1. Two sources were 2 m distant from the microphone array. The distance between the noise and the target was 80 cm. Comparison of the proposed multichannel spatial prediction with conventional spatial prediction for noise reduction is shown in Fig. 2. The

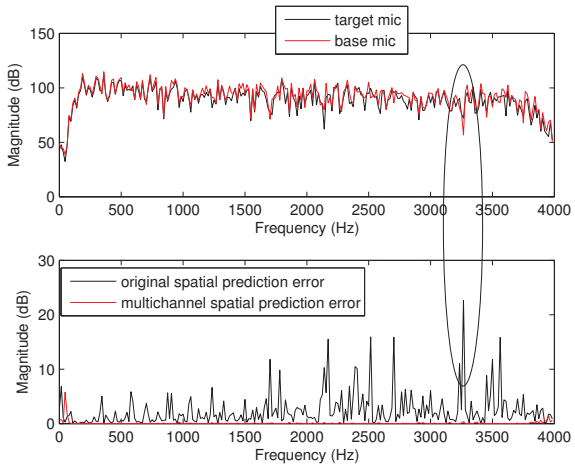


Fig. 2. Comparison of proposed multi-channel spatial prediction with single channel spatial prediction: noise source component in "target mic" is predicted by noise source component in "base mic".

magnitude response is shown in the upper row. In the lower row, the prediction error of each frequency is shown. By multichannel spatial prediction, the prediction error is shown to be remarkably small. Comparison of the proposed method with SP-MDBF [4] is executed. The noise reduction filter length and the distortion-restoration filter length are set to be 8 at each subband. Evaluation measures are PESQ [8], and NRR (Noise reduction Rate), SDR (Signal Distortion Rate), and SIR (Signal Interference Rate). NRR, SDR, and SIR are defined as

$$NRR = 10\log_{10}\frac{\sum_{t=0}^{L}(x_c(t) - y_c(t))^2}{\sum_{t=0}^{L_{wave}}(\hat{y}_c(t) - y_c(t))^2}, \quad (5)$$

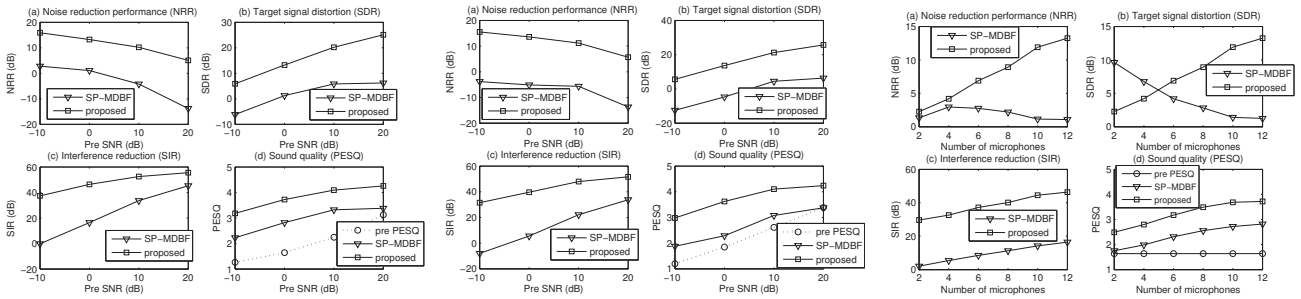
$$SDR = 10\log_{10}\frac{\sum_{t=0}^{L_{wave}}(y_c(t))^2}{\sum_{t=0}^{L_{wave}}(d_c(t))^2}, \quad (6)$$

$$SIR = 10\log_{10}\frac{\sum_{t=0}^{L_{wave}}(y_c(t))^2}{\sum_{t=0}^{L_{wave}}(\hat{n}_c(t))^2}, \quad (7)$$

where $d_c(t)$ is defined as the distortion of the target signal, $\hat{n}_c(t)$ is the residual noise signal. L_{wave} is defined as the length of the wave.

The experimental result for white Gaussian noise is shown in Fig. 3 (A), and the experimental result for human speech noise is shown in Fig. 3 (B). Both experiments are performed by varying SNR of target signal and noise signal. The proposed method outperformed SP-MDBF at all SNR conditions. SIR of the proposed method is higher than 40 dB, noise reduction performance of the proposed method is shown to be extremely high. SDR is higher than 10 dB for SNR $\geq$ 0 dB. The distortion of the output signal is considered to be less than audible level. In SP-MDBF, comparison of white Gaussian noise case with human speech noise case shows that when noise is nonstationary, distortion of the output signal is greatly increased. On the other hand, in the proposed method, when noise is human speech, the degradation of the output signal is less than SP-MDBF. In Fig. 3 (C), experimental results for various number of microphones is shown. The proposed method achieved less distortion signal by increasing the number of microphones.

In Fig. 4, an example of the waves which processed by the proposed method and the conventional method is shown. In the proposed method, the noise reduced signal which is quite similar to the target signal is obtained.



(A) Noise source is white Gaussian. "pre PESQ" means PESQ value of the noisy input signal.

(B) Noise source is speech.

(C) Noise source is white Gaussian. SNR is 0 dB. The number of microphone is changed from 2 to 12.

Fig. 3. Experimental results

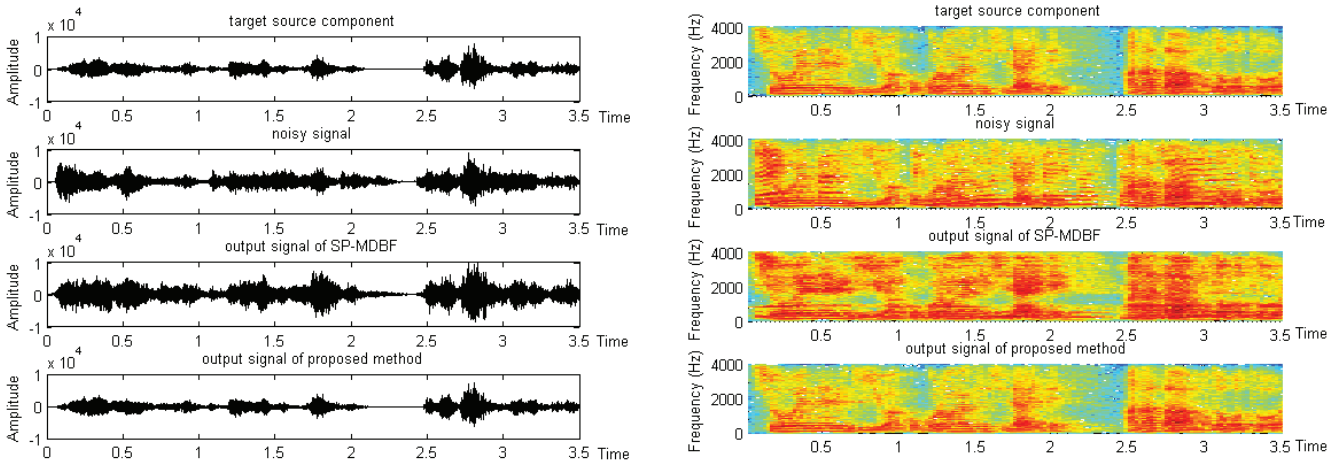


Fig. 4. An example of processed waves by proposed method and SP-MDBF

6. CONCLUSION

In this paper, to achieve good noise reduction performance under reverberant environments, we propose a novel subband noise reduction method based on multichannel spatial prediction. The proposed method is composed of the noise reduction process and the distortion-restoration process. On the noise reduction process, the noise component is reduced without the impulse response of the target source, but the target source component is distorted. The distortion is removed by the distortion-restoration process. The experimental results under the reverberant environment ($RT_{60} \approx 300$ ms) show that the proposed method can reduce more noise than the conventional method and the distortion of the target source is not so big. Complete noise reduction with no distortion of the target source by the proposed framework is assured by MINT theory.

7. REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," In *Proc. IEEE*, vol.60, no.8, pp.926-935, 1972.

[2] L. J. Griffith and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. AP*, vol. 30, i. 1, pp. 27-34, 1982.

[3] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. ASLP*, vol. 15, pp. 1053-1065, 2007.

[4] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. ASLP*, vol. 16, pp. 481-493, 2008.

[5] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. ASSP*, vol. 30, no. 2, pp. 145-152, 1988.

[6] R. Crochiere and L. Rabiner, Multirate Digital Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, *Speech Distortion Weighted Multichannel Wiener Filtering Techniques for Noise Reduction*, Springer-Verlag, ch. 2 in "Speech Enhancement", pp. 199-228, 2005.

[8] ITU-T Rec. P. 862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Geneva, Switzerland, 2001.