Multichannel parametric speech enhancement

DOI:
10.1109/LSP.2005.863819

Document status and date:
Published: 01/01/2006

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Multichannel Parametric Speech Enhancement

Sriram Srinivasan, *Student Member, IEEE*, Robert Aichner, *Student Member, IEEE*, W. Bastiaan Kleijn, *Fellow, IEEE*, and Walter Kellermann, *Member, IEEE*

Abstract—**We present a parametric model-based multichannel approach for speech enhancement. By employing an autoregressive model for the speech signal and using a trained codebook of speech linear predictive coefficients, minimum mean square error estimation of the speech signal is performed. By explicitly accounting for steering errors in the signal model, robust estimates are obtained. Experiments show that the proposed method results in significant performance gains.**

Index Terms—**Acoustic noise, acoustic signal processing, array signal processing, autoregressive processes, speech enhancement, speech processing.**

I. Introduction

In many practical situations that involve the acquisition of speech, the signal is observed in the presence of acoustic background noise. Single-channel noise reduction systems that have only one microphone to record the noisy signal have been popular, especially in mobile communications due to cost and size factors. In applications where multiple microphones provide multiple noisy observations, the spatial diversity can be exploited to achieve noise reduction. A number of multichannel approaches have been proposed, e.g., adaptive beamforming such as the generalized sidelobe canceller [2] and fixed beamformers combined with adaptive post-filters for further noise reduction [3], [4]. While the methods in [3] and [4] assume uncorrelated noise at the different sensors, modifications to deal with diffuse noise were proposed in [5] and [6].

In this letter, we describe a parametric multichannel speech enhancement scheme. By using a trained codebook of clean-speech linear predictive coefficients to parameterize the speech spectrum, a minimum mean square error (MMSE) estimate of the clean speech signal is obtained. A signal model that accounts for microphone array steering errors and diffuse background noise is developed. The resulting multichannel estimator is shown to provide superior performance compared to a recent approach proposed in [6].

II. Signal Model

We assume a far-field model so that wave propagation is planar. The signals arriving at the different sensors differ only in their phase (they are delayed versions of one another). The different sensor signals can be assumed to have identical power spectra, since, in practice, the delay between the sensors is small compared to the short-time stationarity of the speech signal. We assume that the sensor array has been steered toward the direction of the speech source. To ensure a practical model, we allow for steering errors. The additive noise model can then be written as

$$y_i(n) = x_i(n) + w_i(n), \quad 1 \leq i \leq M \tag{1}$$

where $x_i(n)$ is the clean-speech signal component sampled at time instant n at the ith sensor, $w_i(n)$ is the additive noise at the ith sensor, $y_i(n)$ is the noisy speech observed at the ith sensor, and M is the number of sensors. Processing is performed on segments of time-domain samples of length N. In the absence of steering errors, we would have $x_i(n) = x(n)$ for all i. Steering errors are modeled by allowing the clean speech component to differ at the different sensors.

Applying the discrete short-time Fourier transform to a segment of length N, we get

$$Y_i(k) = X_i(k) + W_i(k), \quad 1 \leq i \leq M \tag{2}$$

where $X_i(k) = \sum_{n=0}^{N-1} x_i(n) \exp(-j(2\pi k/N)n)$, and k is the discrete frequency index. $Y_i(k)$ and $W_i(k)$ are similarly obtained from their time-domain counterparts. We define the vectors

$$\begin{aligned} \mathbf{X}(k) &= [X_1(k) \ldots X_M(k)]^T \\ \mathbf{W}(k) &= [W_1(k) \ldots W_M(k)]^T \\ \mathbf{Y}(k) &= [Y_1(k) \ldots Y_M(k)]^T. \end{aligned} \tag{3}$$

Let $\mathbf{P_{xx}}(k) = \mathrm{E}\{\mathbf{X}(k)\mathbf{X}^H(k)\}$ and $\mathbf{P_{ww}}(k) = \mathrm{E}\{\mathbf{W}(k)\mathbf{W}^H(k)\}$ be the $M \times M$ cross spectral density matrices corresponding to the M speech and noise signals. The (i,j)th entry of $\mathbf{P_{xx}}(k)$ and $\mathbf{P_{ww}}(k)$ are given by $\mathrm{E}\{X_i(k)X_j^*(k)\} = P_{x_ix_j}(k)$ and $\mathrm{E}\{W_i(k)W_j^*(k)\} = P_{w_iw_j}(k)$, respectively. Based on our signal model, we have $P_{x_ix_i}(k) = P_{xx}(k)$ and $P_{w_iw_i}(k) = P_{ww}(k)$ for all i, which is justified if the sensors are closely spaced in a homogeneous noise field. Using the definition of the spatial coherence function, the cross spectral densities can be written as

$$\begin{aligned} P_{x_ix_j}(k) &= \Gamma_x^{ij}(k)\sqrt{P_{x_ix_i}(k)P_{x_jx_j}(k)} = \Gamma_x^{ij}(k)P_{xx}(k) \\ P_{w_iw_j}(k) &= \Gamma_w^{ij}(k)\sqrt{P_{w_iw_i}(k)P_{w_jw_j}(k)} = \Gamma_w^{ij}(k)P_{ww}(k) \end{aligned} \tag{4}$$

where $\Gamma_x^{ij}(k)$ is the coherence function corresponding to the speech signals at the ith and jth sensors, and $\Gamma_w^{ij}(k)$ is the coherence function corresponding to the noise signals at these sensors.

Manuscript received August 3, 2005; revised November 1, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dominic K. C. Ho.

S. Srinivasan and W. B. Kleijn are with the Department of Signals, Sensors, and Systems, KTH (Royal Institute of Technology), Stockholm S-100 44, Sweden (e-mail: sriram.srinivasan@s3.kth.se; bastiaan.kleijn@s3.kth.se).

R. Aichner and W. Kellermann are with the University of Erlangen-Nuremberg, Erlangen D-91058, Germany (e-mail: aichner@lnt.de; wk@lnt.de).

Digital Object Identifier 10.1109/LSP.2005.863819

III. Multichannel Parametric Speech Enhancement

Speech is commonly modeled as an autoregressive (AR) process written as

$$x(n) = \sum_{l=1}^{p} a_{x_l} x(n-l) + e(n) \tag{5}$$

where $a_{x_1}, \ldots, a_{x_p}$ are the linear predictive (LP) coefficients of order p, and $e(n)$ is the prediction error, also referred to as the excitation signal. We model $e(n)$ as a Gaussian random process. The LP analysis is typically performed for each short-time frame, within which speech can be assumed to be stationary. Let $\mathbf{y}_i = [y_i(0), \ldots, y_i(N-1)]^T$ denote a frame of length N of the noisy signal observed at sensor i. For each frame, the speech model parameters are the vector of LP coefficients $\theta_x = [1\ a_{x_1}, \ldots, a_{x_p}]$ and the variance of the excitation signal σ_x^2. Given θ_x and σ_x^2, the speech power spectrum $P_{xx}(k)$ is obtained as

$$P_{xx}(k) = \frac{\sigma_x^2}{|A_x(k)|^2}, \quad \text{where } A_x(k) = \sum_{l=0}^{p} a_{x_l} e^{-j\frac{2\pi k}{N} l} \tag{6}$$

with $a_{x_0} = 1$. We denote the speech model parameters as $\psi_x = [\theta_x\ \sigma_x^2]$.

Let θ_w and σ_w^2 denote the noise LP coefficients and excitation variance, respectively, so that the noise power spectrum is given by $P_{ww}(k) = \sigma_w^2 / |A_w(k)|^2$, where $A_w(k)$ is defined similarly to $A_x(k)$. In this letter, we assume that $\psi_w = [\theta_w\ \sigma_w^2]$ is known. In practice, it can be estimated using one of the noise estimation algorithms, e.g., [7]. We assume that the noise has a zero-mean Gaussian distribution.

The multichannel Wiener filter, which provides the MMSE estimate of the clean-speech signal, can be factored into a minimum variance distortionless response (MVDR) beamformer and a single-channel postfilter [8]. To obtain an MMSE estimate of the clean-speech signal from the M noisy observations, it is sufficient to first perform MVDR beamforming to obtain a single-channel beamformer output $\mathbf{y}$ and then estimate the clean signal from $\mathbf{y}$. Let $\mathcal{X}$ denote the random variable corresponding to the clean-speech signal. Consider the noisy observations $\mathbf{y}_1, \ldots, \mathbf{y}_M$, where $\mathbf{y}_i = [y_i(0), \ldots, y_i(N-1)]^T$. The MMSE estimate $\hat{\mathbf{x}}$ of $\mathbf{x} = [x(0) \ldots x(N-1)]$ can thus be written as

$$\hat{\mathbf{x}} = \mathrm{E}\{\mathcal{X} | \mathbf{y}_1, \ldots, \mathbf{y}_M\} = \mathrm{E}\{\mathcal{X} | \mathbf{y}\} \tag{7}$$

which reduces to a single-channel estimation from the signal $\mathbf{y}$. In the case of uncorrelated noise, the MVDR beamformer is a simple delay and sum beamformer, and for diffuse noise, we have a superdirective beamformer [9]. In the most general case, allowing for steering errors and diffuse noise, the beamformer output in the frequency domain can be expressed as $Y(k) = \sum_{i=1}^{M} b_i(k) Y_i(k)$, where $b_i(k)$ are the beamformer weights. The power spectrum of $\mathbf{y}$ can be written as

$$\begin{aligned} P_{yy}(k) &= \mathrm{E}[Y_i Y_i^*] = \sum_{i,j} b_i b_j^* P_{x_i x_j} + \sum_{i,j} b_i b_j^* P_{w_i w_j} \\ &= P_{xx} \underbrace{\sum_{i,j} b_i b_j^* \Gamma_x^{ij}}_{\Gamma_x} + P_{ww} \underbrace{\sum_{i,j} b_i b_j^* \Gamma_w^{ij}}_{\Gamma_w} \\ &\triangleq P_{xx}(k)\Gamma_x(k) + P_{ww}(k)\Gamma_w(k) \end{aligned} \tag{8}$$

where we used (2), (4), and the assumption that speech and noise are uncorrelated[1] and omitted the index k in the first two lines for brevity. The term $\Gamma_x(k)$ accounts for steering errors; in the case of ideal steering, we would have $\Gamma_x(k) = 1$ for all k. We note that $P_{xx}(k)$ and $P_{ww}(k)$ are specified by the speech and noise LP parameters according to (6). The expression for the MMSE estimate $\hat{\mathbf{x}}$ can be rewritten as

$$\begin{aligned} \hat{\mathbf{x}} &= \mathrm{E}\{\mathcal{X}|\mathbf{y}\} = \int_{\Psi_x} p(\psi_x|\mathbf{y}) \mathrm{E}\{\mathcal{X}|\mathbf{y}, \psi_x\} d\psi_x \\ &= \int_{\Psi_x} \frac{p(\mathbf{y}|\psi_x)}{p(\mathbf{y})} p(\psi_x) \mathrm{E}\{\mathcal{X}|\mathbf{y}, \psi_x\} d\psi_x \end{aligned} \tag{9}$$

where $\Psi_x = \Theta_x \times \Sigma_x$, Θ_x represents the support-space of the vectors of speech LP coefficients, and Σ_x corresponds to the support-space for the speech excitation variance.

From the Gaussian assumptions on the speech and noise processes, $p(\mathbf{y}|\psi_x)$ has a zero-mean Gaussian pdf with covariance matrix $\mathbf{R}_\mathbf{y} = \mathrm{E}\{\mathbf{y}\mathbf{y}^T\}$. The covariance matrix is parameterized by ψ_x and ψ_w and is Toeplitz. Assuming that the frame length is large, $\mathbf{R}_\mathbf{y}$ can be approximated as a circulant matrix, which is then diagonalized by the Fourier transform [10]. We have

$$\begin{aligned} p(\mathbf{y}|\psi_x) &= \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{R}_\mathbf{y}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}^T \mathbf{R}_\mathbf{y}^{-1} \mathbf{y}\right) \\ &\approx \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{P}_{yy}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}^T F^H \mathbf{P}_{yy}^{-1} F \mathbf{y}\right) \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{k=0}^{N-1} \sqrt{P_{yy}(k)}} \exp\left(-\frac{1}{2}\sum_{k=0}^{N-1} \frac{|Y(k)|^2}{P_{yy}(k)}\right) \end{aligned} \tag{10}$$

where $F = [f_{pq}]$ denotes the discrete Fourier transform matrix whose (p,q)th entry is given by $f_{pq} = (1/\sqrt{N})\exp(-j2\pi pq/N)$, and $\mathbf{P}_{yy}$ is an $N \times N$ diagonal matrix with $P_{yy}(k)$, $0 \le k \le N-1$ given by (8) on the main diagonal. $P_{yy}(k)$, given by (8), is specified by ψ_x and ψ_w.

For a given θ_x and $\mathbf{y}$, as the excitation variance σ_x^2 deviates from its true value $\tilde{\sigma}_x^2$, the likelihood $p(\mathbf{y}|\psi_x)$ decays rapidly from its maximum value [1]. Thus, approximating $p(\mathbf{y}|\psi_x)$ by $p(\mathbf{y}|\psi_x)\ \delta(\sigma_x^2 - \tilde{\sigma}_x^2)$, we can rewrite (9) as

$$\begin{aligned} \hat{\mathbf{x}} &= \int_{\Psi_x} \frac{p(\mathbf{y}|\psi_x)\delta\left(\sigma_x^2 - \tilde{\sigma}_x^2\right)}{p(\mathbf{y})} p(\psi_x) \mathrm{E}\{\mathcal{X}|\mathbf{y}, \psi_x\} d\psi_x \\ &= \int_{\Theta_x} \frac{p\left(\mathbf{y}|\psi_x'\right)}{p(\mathbf{y})} p\left(\psi_x'\right) \mathrm{E}\left\{\mathcal{X}|\mathbf{y}, \psi_x'\right\} d\theta_x \end{aligned} \tag{11}$$

where $\psi_x' = [\theta_x\ \tilde{\sigma}_x^2]$. Note that instead of an integral over the product space $\Theta_x \times \Sigma_x$, we now have only an integral over Θ_x in (11). In practice, since we do not have the true value $\tilde{\sigma}_x^2$ of the excitation variance, we use an estimate $\hat{\tilde{\sigma}}_x^2$. For computational simplicity, using the relation $P_{yy}(k) =$

[1] While this assumption fails in highly reverberant environments, in moderately reverberant environments such as the one used in the experiments (150 ms), the results indicate that this assumption is satisfied.

$\Gamma_x(k)P_{xx}(k) + \Gamma_w(k)P_{ww}(k)$, we approximate $\hat{\tilde{\sigma}}_x^2$ subtractively according to

$$\hat{\tilde{\sigma}}_x^2 = \frac{\sum_{k=0}^{N-1} \max\left(P_{yy}(k) - \Gamma_w(k)P_{ww}(k), 0\right)}{\frac{\sum_{k=0}^{N-1} \Gamma_x(k)}{|A_x(k)|^2}}. \tag{12}$$

Under our Gaussian model assumptions, the expectation $\mathrm{E}\{\mathcal{X}|\mathbf{y}, \psi'_x\}$ is computed from the Wiener filter (applied to $\mathbf{y}$) estimated in the frequency domain as

$$H(k) = \frac{P'_{xx}(k)}{P'_{xx}(k)\Gamma_x(k) + P_{ww}(k)\Gamma_w(k)} \tag{13}$$

where $P'_{xx}(k) = \tilde{\sigma}_x^2/|A_x(k)|^2$ and $P_{ww}(k) = \sigma_w^2/|A_w(k)|^2$. The estimation of Γ_x and Γ_w is addressed in Section IV. In practice, (11) is computed by a numerical integration performed over a trained codebook of speech LP coefficients

$$\hat{\mathbf{x}} = \frac{1}{N_x} \sum_{i=0}^{N_x} \frac{p\left(\mathbf{y}|\psi'_{x,i}\right) p\left(\tilde{\sigma}_{x,i}^2\right)}{p(\mathbf{y})} \mathrm{E}\left\{\mathcal{X}|\mathbf{y}, \psi'_{x,i}\right\} \tag{14}$$

where $\psi'_{x,i} = [\theta_{x,i}\ \tilde{\sigma}_{x,i}^2]$, $\theta_{x,i}$ is the ith speech codebook entry, $\tilde{\sigma}_{x,i}^2$ is the corresponding speech excitation variance, and N_x is the speech codebook size. The normalizing factor $p(\mathbf{y})$ is given by $p(\mathbf{y}) = (1/N_x)\sum_{i=0}^{N_x} p(\mathbf{y}|\psi'_{x,i})p(\tilde{\sigma}_{x,i}^2)$. Since the speech excitation variance is completely determined given the noisy observation ψ_w and the speech spectral shape θ_x, we assume a noninformative prior (uniform) for the variance. The exact bounds of the uniform distribution are irrelevant since they cancel out in the numerator and denominator of (14). From (14), it can be seen that the MMSE estimator of the clean-speech signal is a weighted sum of Wiener filters corresponding to each entry of the speech codebook. The use of prior information results in a Bayesian MMSE estimate of the speech signal.

IV. Estimation of $\Gamma_x(k)$ and $\Gamma_w(k)$

The coherence terms $\Gamma_x(k)$ and $\Gamma_w(k)$ are required to evaluate the conditional pdf (10) and the excitation variance (12). These can be estimated from the observed data using an approach similar to [11].

From the independence assumption and the additive noise model, the power spectrum of the ith sensor signal satisfies $P_{y_iy_i}(k) = P_{xx}(k) + P_{ww}(k)$ cf. (8). After the beamforming, we have $P_{yy}(k) = P_{xx}(k)\Gamma_x(k) + P_{ww}(k)\Gamma_w(k)$. Let $P_{y_iy_i}^{no}(k)$ denote the value of $P_{y_iy_i}(k)$ in the absence of speech and $P_{y_iy_i}^{sp}(k)$ denote its value in the absence of noise. Let $P_{yy}^{no}(k)$ and $P_{yy}^{sp}(k)$ be defined similarly. Then we have $\Gamma_w(k) = P_{yy}^{no}(k)/P_{y_iy_i}^{no}(k)$ and $\Gamma_x(k) = P_{yy}^{sp}(k)/P_{y_iy_i}^{sp}(k)$. $P_{yy}^{no}(k)$ can be estimated by tracking the minima of $P_{yy}(k)$ binwise since speech energy is not present in all frequency bins at all times.

$P_{yy}^{sp}(k)$ can be estimated by tracking the maxima of $P_{yy}(k)$ binwise. In high energy speech regions, the influence of the noise spectrum is negligible in most practical signal-to-noise ratio (SNR) conditions. For improved accuracy, the estimates are averaged across the M sensors.

For the minimum (maximum) tracking to yield good results, the window over which the tracking is performed should be sufficiently large so as to include noise-only regions (high speech-energy regions for maximum tracking).

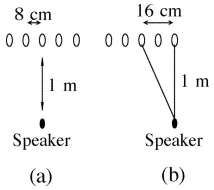

Fig. 1. Microphone array configurations. (a) Ideal steering. (b) A steering error of approximately 9°.

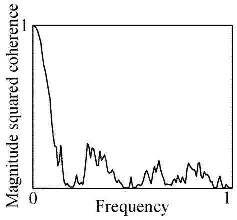

Fig. 2. Plot of the magnitude squared coherence of the babble noise at two sensors, 0.16 m apart. The coherence exhibits characteristics of diffuse noise.

V. Experiments

In this section, we describe the experiments performed to evaluate the proposed multichannel enhancement scheme. A linear microphone array consisting of five equidistant microphones was used in a broadside configuration with an interelement spacing of 8 cm, as shown in Fig. 1. The position of the desired speaker was at a distance of 1 m from the center of the array. The array was placed in a conference room with a reverberation time of approximately 150 ms. Impulse responses from the desired speaker position to each of the five sensors were obtained using the maximum-length sequence approach described in [12]. Clean speech data from the TIMIT database [13] was then convolved with the impulse responses to obtain the speech signals at the sensors. Such a set-up facilitates objective quality measurement since the speech and noise files can be separately processed. Ten seconds each of male and female speech data (8 kHz) were used in the experiments. Office noise (mainly fan noise from four computers) was recorded using the array. We also considered diffuse babble noise, which was artificially generated at the five sensors by positioning 24 different speech sources in a uniformly spaced circular array around the microphones. The coherence function between the noise at sensors spaced 0.16 m apart is shown in Fig. 2.

For the Bayesian approach, a 10-bit codebook of speech LP coefficients of order ten was trained using the Itakura–Saito distortion measure. Ten minutes of speech data (different from the test data) from the TIMIT database were used in the training. The frame length was 240 samples with a 50% overlap, and the frames were Hann windowed. The noise power spectral density was estimated using the minimum statistics technique [7], from which the LP parameters ψ_w were obtained.

For comparison, we also provide results using the postfilter proposed in [6], which is an extension of the Zelinski postfilter [3] to diffuse noise fields. A fixed superdirective beamformer was used for both the proposed Bayesian approach and the reference method. The beamformer weights were computed by using the theoretical expression for the coherence function for diffuse noise fields [9, eq. 2.34].

For evaluation of objective quality, the MVDR beamformer and postfilter computed for the noisy signal were applied to

TABLE I
IMPROVEMENT IN SSNR VALUES WITH AND WITHOUT STEERING ERRORS AT 10 dB INPUT SNR. RESULTS ARE AVERAGED OVER ALL SPEAKERS

	Babble		Office	
Steering Error	0°	9°	0°	9°
Beamformer	2.9	2.4	0.3	-0.3
Reference	3.7	3.6	6.9	6.6
Proposed	6.8	6.7	12.4	11.7

TABLE II
SD VALUES WITH AND WITHOUT STEERING ERRORS AT 10 dB INPUT SNR. RESULTS ARE AVERAGED OVER ALL SPEAKERS

	Babble		Office	
Steering Error	0°	9°	0°	9°
Beamformer	3.2	5.2	3.2	5.2
Reference	4.5	6.4	4.2	6.2
Proposed	4.1	4.5	4.1	4.2

the multichannel clean speech and noise signals at the sensors separately. Using these signals, the segmental SNR (SSNR) at the output was computed. To evaluate the speech distortion, the log-spectral distortion (SD) was computed between the original clean-speech signal (prior to convolving with the impulse responses) and the clean-speech signal at the microphones processed by each of the systems under consideration.

A. Ideal Steering

Table I (columns corresponding to 0°) shows the improvement in SSNR, computed as the difference between the output and input SSNR values for the case of ideal steering [see Fig. 1(a)] for 10-dB input SNR. The reference postfilter performs better than the beamformer, especially for office noise. The office noise has most of its energy in the low-frequency regions where the coherence is high, and thus, the beamformer achieves little improvement. For both babble and office noise, the proposed method results in a significant improvement in performance compared to the reference method.

It can be seen from Table II that under ideal steering (columns corresponding to 0°), the MVDR beamformer has the lowest distortion values, due to its *distortionless* response (the SD value is different from zero as it is computed with respect to the original clean-speech signal, prior to convolving with the impulse responses). As is to be expected, any postprocessing achieves improved noise reduction at the expense of additional signal distortion. From the table, it can be observed that the codebook approach yields lower distortion values than the reference method, while it provides greater noise reduction. Informal listening confirms these improvements.

B. Steering Error of 9°

Both the reference method and the proposed method are robust to mild steering errors in terms of noise reduction, as seen from Table I (columns corresponding to 9°), i.e., the SSNR improvements are similar with and without steering. Again, the proposed method outperforms the reference method by around 3–5 dB.

The methods behave differently in terms of speech distortion (see Table II, columns corresponding to 9°). The beamformer output shows a 2-dB increase in SD due to a 9° steering error. The reference postfilter method also suffers from an increase of almost 2 dB in SD due to the error. In contrast, the proposed method is relatively robust, with only a 0.4-dB increase in SD. It is interesting to note that in the presence of steering errors, using a codebook trained on clean speech compensates for the distortion introduced by the beamformer.

In experiments performed using a larger steering error of 30°, while the overall performance degraded, the proposed method still provided an improvement in SSNR that was around 3 dB higher than the reference schemes. The SD increased by a further 3 dB for all methods.

VI. CONCLUSION

In this letter, we have proposed a parametric MMSE estimation of the clean-speech signal for microphone array speech enhancement. The estimation is performed using an AR model for the clean-speech signal and using a trained codebook of linear predictive coefficients. Using a signal model that incorporates the effect of steering errors, and using a clean-speech codebook, robust performance is achieved, even in the presence of mild steering errors. Experimental results under moderately reverberant conditions and diffuse noise show good performance for the proposed scheme.

ACKNOWLEDGMENT

The authors would like to thank P. Oak for valuable discussions during this work.

REFERENCES

[1] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, Mar. 2005, pp. 1077–1080.

[2] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[3] R. Zelinksi, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, Apr. 1988, pp. 2578–2581.

[4] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electron. Lett.*, vol. 26, no. 24, pp. 2036–2037, Nov. 1990.

[5] R. L. Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 484–487, Sep. 1997.

[6] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[8] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[9] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 2, pp. 19–38.

[10] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 6, pp. 725–730, Nov. 1972.

[11] W. Herbordt, T. Trini, and W. Kellermann, "Robust spatial estimation of the signal-to-interference ratio for nonstationary mixtures," in *Proc. Int. Workshop Acoustic Echo Noise Control*, Sep. 2003, pp. 247–250.

[12] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *J. Audio Eng. Soc.*, vol. 37, no. 6, pp. 419–444, Jun. 1989.

[13] DARPA-TIMIT, in Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, 1990.