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

$$y_k(n) = x_i(n) + w_i(n), \quad 1 \leq i \leq M$$  \hspace{1cm} (1)

where $x_i(n)$ is the clean-speech signal component sampled at time instant $n$ at the $i$th sensor, $w_i(n)$ is the additive noise at the $i$th sensor, $y_k(n)$ is the noisy speech observed at the $i$th 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

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

$$X(k) = [X_1(k) \ldots X_M(k)]^T$$
$$W(k) = [W_1(k) \ldots W_M(k)]^T$$
$$Y(k) = [Y_1(k) \ldots Y_M(k)]^T.$$  \hspace{1cm} (3)

Let $P_{xx}(k) = E\{X(k)X^H(k)\}$ and $P_{ww}(k) = E\{W(k)W^H(k)\}$ be the $M \times M$ cross spectral density matrices corresponding to the $M$ speech and noise signals. The $(i,j)$th entry of $P_{xx}(k)$ and $P_{ww}(k)$ are given by $E\{X_i(k)X_j^*(k)\} = P_{xixj}(k)$ and $E\{W_i(k)W_j^*(k)\} = P_{wiwj}(k)$, respectively. Based on our signal model, we have $P_{xixj}(k) = P_{xx}(k)$ and $P_{wiwj}(k) = P_{ww}(k)$ 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

$$P_{xixj}(k) = \Gamma_{xij}(k) \sqrt{P_{xx}(k)P_{xx}(k)} = \Gamma_{xij}(k)P_{xx}(k)$$
$$P_{wiwj}(k) = \Gamma_{wij}(k) \sqrt{P_{ww}(k)P_{ww}(k)} = \Gamma_{wij}(k)P_{ww}(k)$$  \hspace{1cm} (4)

where $\Gamma_{xij}(k)$ is the coherence function corresponding to the speech signals at the $i$th and $j$th sensors, and $\Gamma_{wij}(k)$ is the coherence function corresponding to the noise signals at these sensors.
III. MULTICHANNEL PARAMETRIC SPEECH ENHANCEMENT

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

\[ x(n) = \sum_{l=1}^{p} a_{x,l} x(n - l) + e(n) \]  

(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 \( y_i = [y_i(0), \ldots, y_i(N - 1)]^T \) 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 variance \( \sigma_x^2 \). Given \( \theta_x \) and \( \sigma_x^2 \), the speech power spectrum \( P_x(k) \) is obtained as

\[ P_x(k) = \frac{\sigma_x^2}{|A_x(k)|^2}, \quad \text{where} \quad A_x(k) = \sum_{l=0}^{p} a_{x,l} e^{-j \frac{2 \pi l k}{N}} \]  

(6)

with \( a_{x,0} = 1 \). We denote the speech model parameters as \( \psi_x = [\theta_x, \sigma_x^2] \).

Let \( \sigma_w^2 \) denote the noise LP coefficients and excitation variance, respectively, so that the noise power spectrum is given by \( P_w(k) = \sigma_w^2/|A_w(k)|^2 \), where \( A_w(k) \) is defined similarly to \( A_x(k) \). In this letter, we assume that \( \sigma_w^2 = \sigma_w^2 \) is known. In practice, it is can be estimated by 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 \( y \) and then estimate the clean signal from \( y \). Let \( \mathcal{X} \) denote the random variable corresponding to the clean-speech signal. Consider the noisy observations \( y_1, \ldots, y_M \), where \( y_i = [y_i(0), \ldots, y_i(N - 1)]^T \). The MMSE estimate \( \hat{x} \) of \( x = [x(0) \ldots x(N - 1)] \) can thus be written as

\[ \hat{x} = \mathbb{E}[\mathcal{X}|y_1, \ldots, y_M] = \mathbb{E}[\mathcal{X}|y] \]  

(7)

which reduces to a single-channel estimation from the signal \( y \).

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 \( y \) can be written as

\[ P_{yy}(k) = \mathbb{E}[Y(k)\mathcal{Y}^H(k)] = \sum_{i,j} b_i b_j^* P_{x,x} + \sum_{i,j} b_i b_j^* P_{w,w} \]  

\[ = P_x \sum_{i,j} b_i b_j^* \Gamma_x^{ij} + P_w \sum_{i,j} b_i b_j^* \Gamma_w^{ij} \]  

\[ = \triangle P_x(k) \Gamma_x(k) + P_w(k) \Gamma_w(k) \]  

(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 \( \hat{x} \) can be rewritten as

\[ \hat{x} = \mathbb{E}[\mathcal{X}|y] = \int \frac{p(\psi_x, y)}{p(y)} \mathbb{E}[\mathcal{X}|y, \psi_x] d\psi_x \]  

\[ \approx \int \frac{p(\psi_x, y)}{p(y)} p(\psi_x) \mathbb{E}[\mathcal{X}|y, \psi_x] d\psi_x \]  

(9)

where \( p(\psi_x, y) = \Theta_x \times \Sigma_x \), \( \Theta_x \) represents the support-space of the vectors of speech LP coefficients, and \( \Sigma_x \) corresponds to the support-space for the speech excitation variance.

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

\[ p(\psi_x, y) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{R}_y|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} y^H \mathbf{R}_y^{-1} y \right) \]  

\[ \approx \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{P}_{yy}(k)|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} y^H \mathbf{F}^H \mathbf{P}_{yy}^{-1} \mathbf{F} y \right) \]  

\[ = \frac{1}{(2\pi)^{\frac{N}{2}} \prod_{l=0}^{N-1} |\mathbf{P}_{yy}(k)|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \sum_{l=0}^{N-1} \mathbf{Y}(k)^H \mathbf{P}_{yy}^{-1} \mathbf{Y}(k) \right) \]  

(10)

where \( \mathbf{F} = \mathbf{F}_{pq} \) denotes the discrete Fourier transform matrix whose \( (p,q) \)th entry is given by \( \mathbf{F}_{pq} = (1/\sqrt{N}) \exp(-j 2\pi pq/N) \), and \( \mathbf{P}_{yy} \) is an \( N \times N \) diagonal matrix with \( \mathbf{P}_{yy}(k), 0 \leq k \leq N - 1 \) given by (8) on the main diagonal. \( \mathbf{P}_{yy}(k), \) given by (8), is specified by \( \psi_x \) and \( \psi_w \).

For a given \( \theta_x \) and \( \theta_w \), as the excitation variance \( \sigma_x^2 \) deviates from its true value \( \sigma_x^2 \), the likelihood \( p(\psi_x|\psi_w) \) decays rapidly from its maximum value [11]. Thus, approximating \( p(\psi_x|\psi_w) \) by \( p(\psi_x|\psi_w) \delta(\sigma_x^2 - \sigma_x^2) \), we can rewrite (9) as

\[ \hat{x} = \int \frac{p(\psi_x, y)}{p(y)} \delta(\sigma_x^2 - \sigma_x^2) p(\psi_x) \mathbb{E}[\mathcal{X}|y, \psi_x] d\psi_x \]  

\[ \approx \int \frac{p(\psi_x, y)}{p(y)} p(\psi_x') \mathbb{E}[\mathcal{X}|y, \psi_x'] d\psi_x' \]  

(11)

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

For computational simplicity, using the relation \( \hat{\sigma}_x^2 = \)\(^1\)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.
\[ \Gamma_x(k)P_{xx}(k) + \Gamma_w(k)P_{uw}(k), \] we approximate \( \hat{\sigma}_x^2 \) subtrac-
tively according to
\[
\hat{\sigma}_x^2 = \sum_{k=0}^{N_x-1} \max \left( P_{yy}(k) - \Gamma_x(k)P_{uw}(k) \right),
\]
(12)

Under our Gaussian model assumptions, the expectation \( E_x \{ x, y, x' \} \) is computed from the Wiener filter (applied to \( y \)) estimated in the frequency domain as
\[
H(k) = \frac{P_{xx}(k)}{P_{xx}(k)\Gamma_x(k) + P_{uw}(k)\Gamma_w(k)}
\]
(13)
where \( P_{xx}(k) = \hat{\sigma}_x^2/|A_x(k)|^2 \) and \( P_{uw}(k) = \hat{\sigma}_w^2/|A_u(k)|^2 \).

Fig. 2. Plot of the magnitude squared coherence of the babble noise at two
speech sources in a uniformly spaced circular array around the
desired speaker (mainly fan noise from four computers) was recorded using the
array. The array was placed in a conference room with a re-
verberation 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 arti-
ficially 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 sens-
ors spaced 0.16 m apart is shown in Fig. 2.

\[ \beta_{x,i} = \theta_{x,i} \hat{\sigma}_{x,i}^2, \]

where \( \hat{\sigma}_{x,i}^2 \) is the \( i \)th speech codebook entry, \( \hat{\sigma}_{x,i}^2 \) is the corresponding speech excitation variance, and \( N_x \) is the speech codebook size. The normalizing factor \( p(y) \) is given by \( p(y) = (1/N_x)\sum_{i=0}^{N_x} p(y) \hat{\sigma}_{x,i}^2 \). Since the speech excitation variance is completely determined given the noisy ob-
servation \( \psi \) and the speech spectral shape \( \theta \), we assume a non-informative 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.

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 micro-
phones was used in a broadband configuration with an interele-
ment 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 re-
verberation 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 arti-
ficially 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 sens-
ors 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 dis-
tortion 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 \( \psi \) 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 refer-
ce 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

<table>
<thead>
<tr>
<th>Steering Error</th>
<th>Babble</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>2.9</td>
<td>3.6</td>
</tr>
<tr>
<td>9°</td>
<td>2.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Reference</td>
<td>3.7</td>
<td>6.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>6.8</td>
<td>12.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Steering Error</th>
<th>Babble</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>3.2</td>
<td>4.5</td>
</tr>
<tr>
<td>9°</td>
<td>5.2</td>
<td>6.4</td>
</tr>
<tr>
<td>Reference</td>
<td>4.5</td>
<td>4.1</td>
</tr>
<tr>
<td>Proposed</td>
<td>4.1</td>
<td>4.2</td>
</tr>
</tbody>
</table>

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