Closely Coupled Array Processing and Model-Based Compensation for Microphone Array Speech Recognition

Xianyu Zhao and Zhijian Ou, *Member, IEEE*

***Abstract*—In conventional microphone array speech recognition, the array processor and the speech recognizer are loosely coupled. The only connection between the two modules is the enhanced target signal output from the array processor, which then gets treated as a single input to the recognizer. In this approach, useful environmental information, which can be provided by the array processor and also needs to be exploited by the recognizer, is ignored. Inherently, the array processor can generate multiple outputs of spatially filtered signals, as a multi-input–multi-output (MIMO) module. In this paper, a closely coupled approach is proposed, in which a recognizer with model-based noise compensation exploits the reference noise outputs from a MIMO array processor. Specifically, a multichannel model-based noise compensation is presented, including the compensation procedure using the vector Taylor series (VTS) expansion and parameter estimation using the expectation-maximization (EM) algorithm. It is also shown how to construct MIMO array processors from conventional beamformers. A number of practical implementations of the conventional loosely coupled approach and the proposed closely coupled approach were tested on a publicly available database, the Multichannel Overlapping Number Corpus (MONC). Experimental results showed that the proposed closely coupled approach significantly improved the speech recognition performance in the overlapping speech situations.**

***Index Terms*—Array signal processing, microphone array, model-based compensation, robust speech recognition.**

I. Introduction

RECENT research efforts in automatic speech recognition (ASR) have been focused on improving the robustness of ASR systems in practical applications, e.g., with spontaneous casual speech, in adverse acoustic conditions, etc. Particularly, in many real environments, such as vehicles, meeting rooms, and information kiosks, the use of hand-held or head-mounted close-talking microphones is undesirable for reasons of safety or convenience. Users expect to speak at some distance from the microphone in a hands-free mode. Unfortunately, in these distant-talking settings, the performance of speech recognizers will be seriously degraded due to the microphone pickup of environmental noise and reverberation. It is known that the use of a microphone array, rather than a single microphone, can suppress interfering signals coming from undesired directions by providing spatial filtering to the sound field [1].

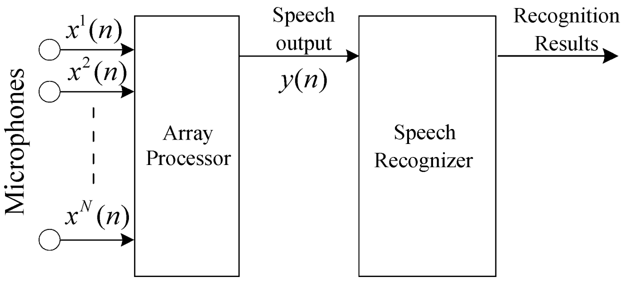

Fig. 1. Loosely coupled array processor and speech recognizer.

Many microphone array processing techniques for interference suppression have been proposed, mainly to enhance the signal-to-noise ratio (SNR) of the target signal. The simplest and most common method is the delay-and-sum (DAS) beamforming [2]. Generalized sidelobe canceller (GSC) is a widely used adaptive beamforming algorithm [3]. To mitigate the undesirable effect of signal cancellation in reverberant environments, the basic GSC can be augmented with coefficient-constrained blocking matrix, adaptive mode control and norm-constrained multiple canceller [4], [5]. Furthermore, the beamformer output can be further enhanced by applying a post-filter [6]–[10]. It is shown [6] that a typical beamformer such as GSC alone does not provide sufficient noise reduction for a broadband input such as speech, and so various post-filtering techniques are proposed [7]–[10] to further filter out the residual noise in the beamformer output.

When used for speech recognition, these microphone array processing methods conventionally take in the multichannel input and generate the enhanced target signal as a single-channel output, which then gets treated as a single-channel input to the recognizer [10]–[12]. The two modules, the array processor and the speech recognizer, are thus loosely coupled, as shown in Fig. 1. The only connection between them is the target signal output from the array processor.

This loosely coupled approach has inherently two problems. The first is the loose coupling of the design objectives of the two modules. The array processor is designed to maximize the

Manuscript received September 5, 2005; revised May 23, 2006 This work was supported by the National Natural Science Foundation of China (No. 60402029). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mari Ostendorf.

X. Zhao was with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. He is now with the France Telecom Research and Development Center, Beijing 100084, China (e-mail: xianyu.zhao@rd.francetelecom.com).

Z. Ou is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: ozj@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TASL.2006.881673

SNR, while the speech recognizer is designed to maximize the likelihood of the acoustic observations under hypothesized word string. To address this mismatch problem, a new approach called likelihood-maximizing beamforming (LIMABEAM) was proposed by Seltzer *et al.* [13]. The beamformer's weights are designed so as to maximize the likelihood of the filtered acoustic data under the recognizer's best hypothesis. Maximum likelihood beamforming was further developed in the cepstral domain by Raub *et al.* [14].

The second is the loose coupling of the operations of the two modules, which is the main issue addressed in this paper. Inherently, the array processor can generate multiple outputs of spatially filtered signals, as a multi-input–multi-output (MIMO) module. However, conventionally the array processor simply operates as a multi-input–single-output (MISO) module without any regard to the manner in which the speech recognizer operates. Useful environmental information that can be provided by the microphone array processor and also needs to be exploited by the recognizer is ignored. We could have the array processor give multiple outputs: one for the enhanced target speech signal, and the others for the estimates of the spatial noise field, which we call the reference noise outputs. The recognizer could then utilize the reference noise outputs in further model-based compensation[1] for robust speech recognition.

Many techniques have been developed, in mono-microphone situations, to augment the basic recognizer with additional components to cope with environmental interference such as additive noise. One important class is model-based noise compensation, e.g., parallel model combination (PMC) [15], vector Taylor series (VTS) [16]–[19], etc. Usually, we first assume an environment model, which describes how the clean speech is corrupted by environmental interference to produce the distorted speech. Based on the environment model, the compensation can then be done in two forms: the model parameters are adapted (model adaptation) [16], [17], or the clean speech is estimated (data compensation) [17]–[19].

One critical step in these model-based compensation methods is how to effectively model the environmental interference such as additive noise. When using only one microphone for speech acquisition, what the recognizer has is only the noisy speech.[2] The additive noise model is usually obtained from the nonspeech frames at the start and/or the end of the noisy utterance. It is also possible to reestimate the noise parameters, after initialization from the nonspeech frames. In these cases [15]–[19], the noise model is used time-invariantly for the whole utterance. However, the noise in real environments is often nonstationary. For example, in meeting environments with several competing speakers, the statistics of the overlapping speech are highly time-varying [20]. Some complicated methods have been proposed for time-varying noise compensation in the case of using only one microphone [21]–[23]. However, microphone arrays can naturally provide estimates of the environmental noise through spatial filtering as pointed above. It is beneficial to closely couple the operations of the array processor and the speech recognizer, with the model-based noise compensation exploiting the noise estimates coming from the MIMO array processor.

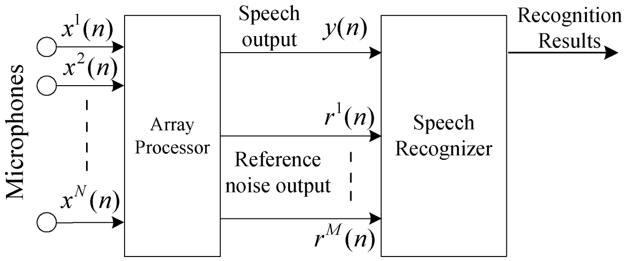

Fig. 2. Closely coupled array processor and speech recognizer.

In this paper, a new approach to microphone array speech recognition is proposed, in which the operations of the array processor and the speech recognizer are closely coupled [24], as shown in Fig. 2. Specifically, we consider the integration of MIMO processing of the microphone array with VTS model-based noise compensation. Here, the array processor provides multiple outputs, allowing more information to be used in the recognizer. One is the enhanced target speech output; the others are the reference noise outputs. The conventional VTS algorithm is extended to utilize the reference noise outputs from the array processor to compensate for the residual noise in the enhanced target speech output. A multichannel environment model is proposed for this compensation purpose. Moreover, an iterative method using the expectation-maximization (EM) algorithm [25] is developed to estimate the compensation parameters. It is also shown how to construct MIMO array processors from conventional beamformers. A number of practical implementations of the conventional loosely coupled approach and the proposed closely coupled approach were tested on the Multichannel Overlapping Numbers Corpus (MONC) database.[3] Experimental results showed that the proposed closely coupled approach significantly improved the speech recognition performance in the overlapping speech situations.

The rest of the paper is organized as follows. In Section II, we focus on the recognizer module for the closely coupled microphone array speech recognition. A multichannel environment model that considers multiple outputs of the MIMO array processor is proposed, and then model-based noise compensation with such multichannel environment model is presented in detail, including the compensation procedure and parameter estimation. In Section III, we discuss the array processor module for the closely coupled approach by showing how to construct MIMO array processors from some conventional beamformers like DAS and GSC. In Section IV, a number of practical implementations of the conventional loosely coupled approach and the proposed closely coupled approach are presented and evaluated through a series of speech recognition experiments on the MONC database. Section V concludes the paper with a summary.

[1] To be more precise, here it is mainly to compensate for the effect of the residual noise, which still exists in the enhanced target speech signal from the beamformer.

[2] And in the loosely coupled approach to using microphone arrays, what the recognizer has is only the single enhanced target speech signal from the beamformer, still with residual noise.

[3] Multi-Channel Overlapping Numbers Corpus (MONC) Distribution. [Online]. Available: http://cslu.cse.ogi.edu/corpora/

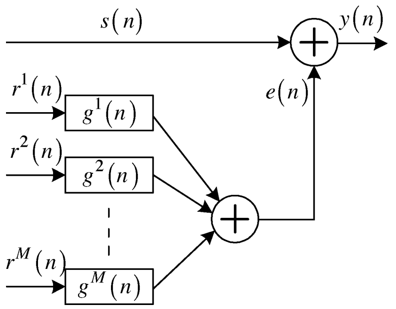

Fig. 3. Multichannel environment model.

II. Closely Coupled Operations for Microphone Array Speech Recognition: Recognizer Module

In this section, we assume that the microphone array processor operates as a MIMO module and describe how the speech recognizer can make use of these multiple outputs for the purpose of model-based noise compensation. (The implementation of MIMO array processor will be presented in Section III.) First, a multichannel environment model is proposed, which takes into account the relationship between the desired clean speech and the array processor's multiple outputs. Then, the model-based noise compensation procedure based on VTS expansion of this multichannel environment model is described in detail. To address the problem of maximum-likelihood (ML) estimation of the set of equalization filters used in the compensation procedure, we develop an iterative method using the EM algorithm.

A. Multichannel Environment Model

When working as a MIMO module as shown in Fig. 2, the microphone array processor takes several channels of signals acquired by the array sensors as inputs. It also has several output channels, which include the enhanced target speech $y(n)$ and the reference noise outputs, $\{r^1(n), r^2(n), \ldots, r^M(n)\}$. Although the SNR of the enhanced target speech $y(n)$ is increased compared with the microphone inputs, there is still some residual noise in it [6]. We formulate the relationship between the desired clean speech signal $s(n)$, the array processor output of enhanced speech $y(n)$, and the residual noise $e(n)$ as follows:

$$y(n) = s(n) + e(n). \tag{1}$$

In addition, we assume that the residual noise $e(n)$ could be represented by a combination of the filtered reference noise outputs, as

$$e(n) = \sum_{j=1}^{M} r^j(n) \otimes g^j(n) \tag{2}$$

where $\{g^1(n), g^2(n), \ldots, g^M(n)\}$ stands for a set of equalization filters, which account for the amplitude and phase differences between $e(n)$ and $r^j(n)$'s. Combining (1) and (2), the environment model considering the array processor's multichannel outputs is shown in Fig. 3.

The aforementioned relationship can be represented using power spectral density (PSD) as

$$|Y(\omega)|^2 = |S(\omega)|^2 + \sum_{j=1}^{M} \left|R^j(\omega)\right|^2 \cdot \left|G^j(\omega)\right|^2 \tag{3}$$

where $|Y(\omega)|^2$, $|S(\omega)|^2$, $|R^j(\omega)|^2$, and $|G^j(\omega)|^2$ are the PSD of $y(n)$, $s(n)$, $r^j(n)$, and $g^j(n)$, respectively, and for simplicity, we assume that $s(n)$ and $\{r^j(n);\ j = 1, \ldots, M\}$ are mutually independent.[4] After taking natural logarithm on (3), we get the relationship in the logarithm filter bank energy (log-FBE) domain

$$\log|Y(\omega)|^2 = \log|S(\omega)|^2 + \log\left[1 + \sum_{j=1}^{M} \left|R^j(\omega)\right|^2 \cdot \left|G^j(\omega)\right|^2 / |S(\omega)|^2\right]. \tag{4}$$

For brevity, we use y_l, s_l, $\{r_l^j;\ j = 1, \ldots, M\}$, and $\{g_l^j;\ j = 1, \ldots, M\}$ to represent $\log|Y(\omega)|^2$, $\log|S(\omega)|^2$, $\{\log|R^j(\omega)|^2;\ j = 1, \ldots, M\}$, and $\{\log|G^j(\omega)|^2;\ j = 1, \ldots, M\}$, respectively, where the subscript "l" denotes log-FBE domain. After some algebraic manipulation, (4) can be rewritten with these new symbols as

$$y_l = s_l + \log\left(1 + \sum_{j=1}^{M} \exp\left(r_l^j + g_l^j - s_l\right)\right). \tag{5}$$

B. Multichannel Model-Based Noise Compensation With VTS

The clean speech signal s_l is modeled by a K-Gaussian mixture in the log-FBE domain as

$$p(s_l) = \sum_{k=1}^{K} p(v = k) N(s_l; \mu_{s,k}, \Sigma_{s,k}) \tag{6}$$

where $p(v = k)$ is the *a priori* probability of the kth Gaussian component which has mean $\mu_{s,k}$ and covariance matrix $\Sigma_{s,k}$.

We treat the equalization filters $\{g_l^j;\ j = 1, \ldots, M\}$ as unknown constant parameters rather than random variables, because their values change more slowly compared with those of speech and noise signals.

To facilitate model-based noise compensation, the nonlinear model (5) can be approximated with its first-order vector Taylor series expansion. For the kth Gaussian component in the clean speech model, by treating $\{r_l^j;\ j = 1, \ldots, M\}$ as observations and expanding VTS around the mean of the clean speech $(\mu_{s,k})$ and the current values of equalization filters $(\bar{g}_l^j)$, $j = 1, \ldots, M$, i.e., $\{\mu_{s,k}, \bar{g}_l^j;\ j = 1, \ldots, M\}$, we get

$$y_l \approx \mu_{s,k} + \log\left(1 + \sum_{j=1}^{M} \exp\left(r_l^j + \bar{g}_l^j - \mu_{s,k}\right)\right) + A_k(s_l - \mu_{s,k}) + \sum_{j=1}^{M} B_{j,k}\left(g_l^j - \bar{g}_l^j\right) \tag{7}$$

[4] This will not be the real case, since leakage will cause that the enhanced target signal and the reference noise outputs are correlated.

where (diag is an operator of forming a diagonal matrix from a vector)

$$
\begin{aligned}
A_k &= \left.\frac{\partial y_l}{\partial s_l}\right|_{\{\mu_{s,k},\overline{g}_l^j\}}\\
&= \mathrm{diag}\left(1+\sum_{j=1}^{M}\exp\left(r_l^j+\overline{g}_l^j-\mu_{s,k}\right)\right)^{-1}\\
B_{j,k} &= \left.\frac{\partial y_l}{\partial g_l^j}\right|_{\{\mu_{s,k},\overline{g}_l^j\}}\\
&= \mathrm{diag}\left(\frac{\exp\left(r_l^j+\overline{g}_l^j-\mu_{s,k}\right)}{1+\sum\limits_{j'=1}^{M}\exp\left(r_l^{j'}+\overline{g}_l^{j'}-\mu_{s,k}\right)}\right).
\end{aligned}
$$

Model-based compensation is done to the mean vector and covariance matrix for each Gaussian component in the clean speech model. With the modeling considerations from (5)–(7), the noisy signal y_l is also modeled as a Gaussian mixture, and its mean and covariance matrix for the kth component can be shown as

$$
\mu_{y,k} = \mu_{s,k} + \log\left(1+\sum_{j=1}^{M}\exp\left(r_l^j+\overline{g}_l^j-\mu_{s,k}\right)\right) \tag{8}
$$

and

$$
\Sigma_{y,k} = A_k\Sigma_{s,k}A_k^T. \tag{9}
$$

Since the interferences in real environments are often nonstationary (e.g., in the case of overlapping speech), the model compensation is carried out frame by frame as follows. From now on, we include frame t in parentheses to explicitly express the dependence on frame index.

$$
\mu_{y,k}(t) = \mu_{s,k} + \log\left(1+\sum_{j=1}^{M}\exp\left(r_l^j(t)+\overline{g}_l^j-\mu_{s,k}\right)\right) \tag{10}
$$

and

$$
\Sigma_{y,k}(t) = A_k(t)\Sigma_{s,k}A_k^T(t) \tag{11}
$$

where

$$
\begin{aligned}
A_k(t) &= \left.\frac{\partial y_l}{\partial s_l}\right|_{\{\mu_{s,k},\overline{g}_l^j\}}\\
&= \mathrm{diag}\left(1+\sum_{j=1}^{M}\exp\left(r_l^j(t)+\overline{g}_l^j-\mu_{s,k}\right)\right)^{-1}\\
B_{j,k}(t) &= \left.\frac{\partial y_l}{\partial g_l^j}\right|_{\{\mu_{s,k},\overline{g}_l^j\}}\\
&= \mathrm{diag}\left(\frac{\exp\left(r_l^j(t)+\overline{g}_l^j-\mu_{s,k}\right)}{1+\sum\limits_{j'=1}^{M}\exp\left(r_l^{j'}(t)+\overline{g}_l^{j'}-\mu_{s,k}\right)}\right).
\end{aligned}
$$

After model compensation, the clean speech is estimated based on the minimum mean square error (MMSE) criterion. Thus, we have

$$
\begin{aligned}
\hat{s}_l(t) &= E\left[s_l(t)|y_l(t),\left\{r_l^j(t);j=1,\ldots,M\right\}\right]\\
&= \sum_{k=1}^{K}P\left(v(t)=k|y_l(t),\left\{r_l^j(t);j=1,\ldots,M\right\}\right)\hat{s}_l(t,k)
\end{aligned} \tag{12}
$$

where

$$
\hat{s}_l(t,k) = \mu_{s,k} + A_k^{-1}(t)\cdot\left(y_l(t)-\mu_{y,k}(t)\right).
$$

Using the compensated model (10) and (11) for $y_l(t)$, the *posteriori* probabilities $P(v(t)=k|y_l(t),\{r_l^j(t);\ j=1,\ldots,M\})$ are computed as

$$
\begin{aligned}
&P\left(v(t)=k|y_l(t),\left\{r_l^j(t);j=1,\ldots,M\right\}\right)\\
&\quad= \frac{P(v=k)N\left(y_l(t);\mu_{y,k}(t),\Sigma_{y,k}(t)\right)}{\sum\limits_{k'=1}^{K}P(v=k')N\left(y_l(t);\mu_{y,k'}(t),\Sigma_{y,k'}(t)\right)}.
\end{aligned} \tag{13}
$$

C. Maximum Likelihood Estimation of the Equalization Filters

The estimation of the equalization filters, $\{g_l^j;\ j=1,\ldots,M\}$, between the residual noise and the reference noises can be based on the maximum-likelihood criterion. Since it is difficult to obtain the ML estimate directly, the EM algorithm is used to iteratively update the parameter values. The auxiliary function $Q(\lambda|\overline{\lambda})$ for the EM algorithm is defined as follows, for current parameters $\overline{\lambda}=\{\overline{g}_l^j;\ j=1,\ldots,M\}$ and the parameters $\lambda=\{g_l^j;\ j=1,\ldots,M\}$ to be reestimated

$$
Q(\lambda|\overline{\lambda}) = E\left[\log p(Y_l,R_l,S_l,V|\lambda)|Y_l,R_l,\overline{\lambda}\right] \tag{14}
$$

where $Y_l=y_l(1),\ldots,y_l(T)$ is the noisy feature vector sequence of length T, $S_l=s_l(1),\ldots,s_l(T)$ is the clean feature vector sequence, $R_l=\{r_l^j(t);\ j=1,\ldots,M$ and $t=1,\ldots,T\}$ are the reference noise feature vectors and $V=v(1),\ldots,v(T)$ is the hidden sequence of mixture components. The parameter reestimate $\hat{\lambda}$ is obtained through the following optimization problem with respect to λ, i.e.,

$$
\hat{\lambda} = \arg\max_{\lambda} Q(\lambda|\overline{\lambda}). \tag{15}
$$

Expanding (14), we have

$$
\begin{aligned}
&Q(\lambda|\overline{\lambda})\\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}P\left(v(t)=k|y_l(t),\left\{r_l^j(t);j=1,\ldots,M\right\},\overline{\lambda}\right)\\
&\quad\cdot\log\left[p\left(y_l(t)|v(t)=k,\left\{r_l^j(t);j=1,\ldots,M\right\},\lambda\right)\right]+C_1\\
&= -\frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(v(t)=k|y_l(t),\left\{r_l^j(t);j=1,\ldots,M\right\},\overline{\lambda}\right)\\
&\quad\times\left(h(t,k)\right)^T\Sigma_{s,k}^{-1}\left(h(t,k)\right)+C_2
\end{aligned} \tag{16}
$$

where

$$
h(t,k) = A_k^{-1}(t)\left(y_l(t)-\mu_{y,k}(t)-\sum_{j=1}^{M}B_{j,k}(t)\left(g_l^j-\overline{g}_l^j\right)\right).
$$

C_1 and C_2 are constant values that are not relevant to the optimization problem.

To maximize $Q(\lambda|\bar{\lambda})$, we let $(\partial Q(\lambda|\bar{\lambda})/\partial g_l^m) = 0, m = 1,\ldots,M$ and obtain for $m = 1,\ldots,M$

$$\begin{aligned}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{M} & P\left(v(t) = k|y_l(t), \left\{r_l^j(t); j = 1,\ldots,M\right\}, \overline{\lambda}\right) \\ & \times B_{m,k}^T(t)A_k^{-1}(t)\Sigma_{s,k}^{-1}A_k^{-1}(t)B_{j,k}(t)\left(g_l^j - \overline{g}_l^j\right) \\ & = \sum_{t=1}^{T}\sum_{k=1}^{K} P\left(v(t) = k|y_l(t), \left\{r_l^j(t); j = 1,\ldots,M\right\}, \overline{\lambda}\right) \\ & \quad \times B_{m,k}^T(t)A_k^{-1}(t)\Sigma_{s,k}^{-1}A_k^{-1}(t)\left(y_l(t) - \mu_{y,k}(t)\right). \qquad (17)\end{aligned}$$

To be more precise, the new estimates of the M equalization filters $(\hat{g}_l^1,\ldots,\hat{g}_l^M)$ are found by solving the following vector linear equations:

$$W \cdot \hat{g} = z \qquad (18)$$

where

$$W=\begin{pmatrix} w_{11} & \cdots & w_{1M} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & w_{MM}\end{pmatrix} \quad \hat{g}=\begin{pmatrix} \hat{g}_l^1 - \overline{g}_l^1 \\ \vdots \\ \hat{g}_l^M - \overline{g}_l^M\end{pmatrix} \quad z=\begin{pmatrix} z_1 \\ \vdots \\ z_M\end{pmatrix}$$

and

$$\begin{aligned} w_{mj} &= \sum_{t=1}^{T}\sum_{k=1}^{K} P\left(v(t)=k|y_l(t), \left\{r_l^j(t); j=1,\ldots,M\right\}, \overline{\lambda}\right) \\ &\quad \times B_{m,k}^T(t)A_k^{-1}(t)\Sigma_{s,k}^{-1}A_k^{-1}(t)B_{j,k}(t) \\ z_m &= \sum_{t=1}^{T}\sum_{k=1}^{K} P\left(v(t)=k|y_l(t), \left\{r_l^j(t); j=1,\ldots,M\right\}, \overline{\lambda}\right) \\ &\quad \times B_{m,k}^T(t)A_k^{-1}(t)\Sigma_{s,k}^{-1}A_k^{-1}(t)\left(y_l(t) - \mu_{y,k}(t)\right).\end{aligned}$$

The new values of the equalization filters $\hat{g}_l^j$ obtained above are then used in the next round of EM iteration as the current estimates $\overline{g}_l^j$.

Before the EM iterations, the initial values of g_l^j can be set according to the amplitude ratio estimate between the residual noise and the reference noises. These can be estimated from the nonspeech frames in the array's target output, where the voice activity detector (VAD) marks as not being spoken by the target speaker. Then we have

$$\overline{g}_l^j(b)=\log\left[\left\langle|Y(b)|^2\right\rangle\Big/\sum_{j'=1}^{M}\left\langle\left|R^{j'}(b)\right|^2\right\rangle\right], \quad j=1,2,\ldots,M \qquad (19)$$

where b is the filter bank index, and the symbol "$\langle\ \rangle$" represents averaging over the nonspeech frames.

In conclusion, the multichannel model-based noise compensation proceeds as follows [17], [18].

1) Get initial estimates for g_l^j using (19).
2) Update the compensated model (10), (11) for the noisy signal $y_l(t)$, using the current estimates of the equalization filters.
3) Perform a single iteration of the EM algorithm to reestimate the equalization filters.
4) If the likelihood of the noisy signal $y_l(t)$ increases relatively above a predefined ratio (e.g., 0.1%), or the current number of iterations is below a fixed number (e.g., 10),[5] return to Step 2.
5) The clean speech is estimated based on the MMSE criterion through (12).

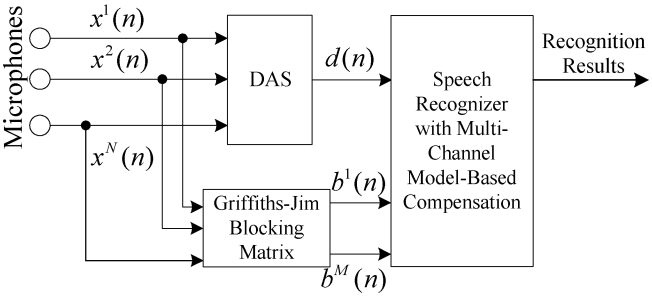
Microphones
$x^1(n)$
$x^2(n)$
$x^N(n)$
DAS
$d(n)$
Speech Recognizer with Multi-Channel Model-Based Compensation
Recognition Results
Griffiths-Jim Blocking Matrix
$b^1(n)$
$b^M(n)$

Fig. 4. Closely coupled DAS-BM MIMO array processor and multichannel model-based noise compensation.

III. Closely Coupled Operations for Microphone Array Speech Recognition: Array Processor Module

In Section II, it is assumed that we have a MIMO array processor module, which has several output channels, including not only the enhanced target speech $y(n)$, but also some reference noise outputs, $\{r^1(n), r^2(n),\ldots,r^M(n)\}$. In this section, we detail how to construct MIMO array processors in practice, from some conventional beamformers like DAS and GSC. Two different schemes are presented.

A. DAS-BM MIMO Array Processor

The DAS-BM MIMO array processor includes a delay-and-sum (DAS) beamformer and a blocking matrix (BM), as shown in Fig. 4.

DAS is used to get steered response from the multisensor array for the target direction. It just applies time shifts to the array signals, $\{x^i(n);\ i = 1,\ldots,N\}$, to compensate for the propagation delays, $\{\tau^i;\ i = 1,\ldots,N\}$, in the arrival of the target signal at the microphones. The array signals are time-aligned and summed together to form a single output signal $d(n)$, i.e.,

$$d(n) = \frac{1}{N}\sum_{i=1}^{N} x^i(n + \tau^i). \qquad (20)$$

In our experiments, the steering delays for the target speaker are estimated using the PHAT method [26].

BM is used to block the desired speech signal and passes the speech from other competing speakers. So in the BM outputs, $\{b^j(n);\ j = 1,2,\ldots,M\}$, the interfering signals are dominant. In our experiments, the Griffiths–Jim blocking matrix [3] is used as the BM. It simply takes the difference between the adjacent time aligned array signals to get the $b^j(n)$, i.e.,

$$b^j(n) = x^j(n + \tau^j) - x^{(j+1)}\left(n + \tau^{(j+1)}\right). \qquad (21)$$

[5] Note that since the linear expansion (of the nonlinear environment model) changes from iteration to iteration with changing estimates of the equalization filters, the objective function being maximized by the EM algorithm changes from iteration to iteration. As such, there is no guarantee of likelihood convergence. Thanks for the comment from reviewer 1.

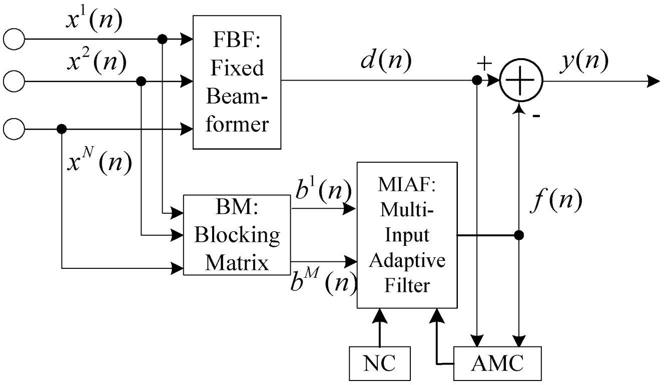

Fig. 5. Robust generalized sidelobe canceller incorporating adaptive mode control (AMC) and norm constraint (NC).

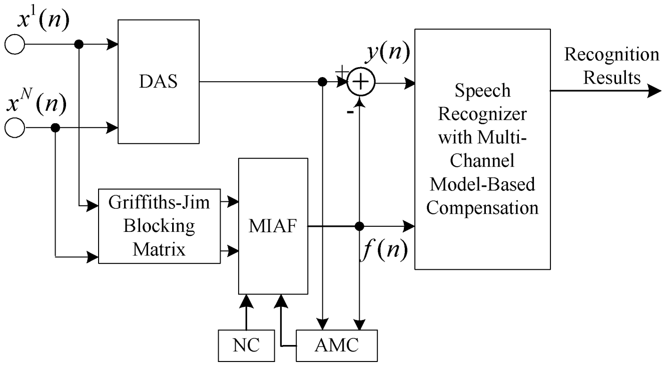
$x^1(n)$
$x^N(n)$
DAS
$y(n)$
+
-
Speech Recognizer with Multi-Channel Model-Based Compensation
Recognition Results
Griffiths-Jim Blocking Matrix
MIAF
$f(n)$
NC
AMC

Fig. 6. Closely coupled robust GSC MIMO array processor and multichannel model-based noise compensation.

For the DAS-BM MIMO processor, the closely coupled approach uses the BM outputs as the reference noise outputs (i.e., $r^j(n) = b^j(n), j = 1, 2, \ldots, M$) and does multichannel model-based noise compensation for the residual noise in the DAS beamformer output.

B. GSC MIMO Array Processor

Generalized sidelobe canceller (GSC) is a more sophisticated beamformer. It attempts to suppress the noise by constraining the array response to unity in the direction of the target speech and minimizing the energy from all other directions. It includes a fixed beamformer (FBF), a multi-input adaptive filter (MIAF) and a blocking matrix (BM). In our experiments, the FBF and BM are implemented as the DAS beamformer and the Griffiths–Jim blocking matrix, respectively, as described previously. The MIAF uses the normalized least mean square (NLMS) algorithm [27] to adapt the coefficients of a set of transversal FIR filterers to meet the minimum variance distortionless response (MVDR) criterion. The desired speech output $y(n)$ is obtained by subtracting the output of MIAF $f(n)$ from the FBF output $d(n)$.

Although the convergence properties of the GSC algorithm have been shown in [2], [3], the real environments for microphone array applications are more complicated. Factors such as room reverberation and nonstationary interference will cause the leakage of target speech signal into the BM outputs, the target signal cancellation during the subtraction of FBF and MIAF outputs, and the unstable adaptation of MIAF coefficients [5]. All these deteriorate the performance for target speech enhancement, and so for subsequent speech recognition [13]. In [5], Hoshuyama discusses several robust adaptive beamforming techniques, like adaptive mode control (AMC), norm constraint (NC) of MIAF coefficients, etc. With AMC, the coefficients of MIAF are allowed to adapt only when the noises are dominant in the BM outputs, and NC is used to constrain the amplitude growth of these coefficients to further avoid incorrect and unstable adaptation. A robust GSC used in our experiments that incorporates AMC and NC is shown in Fig. 5.

For the GSC MIMO processor, the closely coupled approach uses the MIAF output as the reference noise output (i.e., $M = 1$ and $r^1(n) = f(n)$) and does multichannel model-based noise compensation for the residual noise in the GSC beamformer output, as shown in Fig. 6.

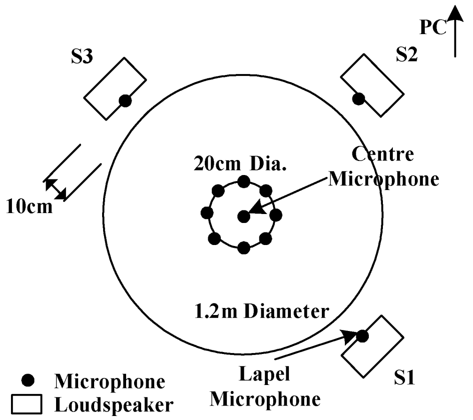

Fig. 7. Meeting room configuration for MONC .

IV. Experimental Results

In Sections II and III, we introduce the recognizer module and the array processor module, respectively, for the proposed closely coupled approach to microphone array speech recognition. In order to evaluate the proposed approach, we employed a publicly available database, the Multichannel Overlapping Numbers Corpus (MONC). The MONC is based on the Numbers Corpus (telephone quality, continuous sentence with 30-word vocabulary) prepared by the Center for Spoken Language Understanding at the Oregon Graduate Institute . The meeting room configuration for the MONC data acquisition is shown in Fig. 7. The loudspeakers simulate the presence of the desired speaker (S1) and the two competing speakers (S2 and S3) in a realistic meeting scenario. A circular microphone array comprising eight equally spaced microphones is placed in the middle of a round table ($N = 8$). An additional microphone is placed at the center of the table. A lapel microphone is attached to each loudspeaker. The same type omnidirectional microphones are used in all locations. The circular table is located at one end of a moderately reverberant, 8.2 × 3.6 × 2.4 m, rectangular room. The dominant nonspeech noise is produced by a PC located at the opposite end of the room. There are three possible competing speaker scenarios.

- Scenario-S1: Only the desired speaker S1 is active, no overlapping speech.

- Scenario-S1S2: The desired speaker S1 with one competing speaker S2 active (resulting in approximately 0 dB SNR at the center table-top microphone location).
- Scenario-S1S2S3: The desired speaker S1 with two competing speakers S2 and S3 active (resulting in approximately -3 dB SNR at the center location).

The speech recognition system is based on continuous-density hidden Markov models (HMM) with 363 states. Each word is modeled by a left-to-right 12-state HMM. The silence is modeled as a three-state HMM. The observation probability distribution for each state is a six-Gaussians mixture. The 45-dimensional feature vector is formed by 14 MFCCs, the energy plus their first and second-order differentials. A 20-ms window length and a 10-ms frame shift are used. Cepstral mean normalization (CMN) is performed in both training and test.

The clean corpus is comprised of a 6049-sentence training set, a 2026-sentence cross-validation set, and a 2061-sentence test set. Different "scenario" versions of the cross-validation and test sets were collected by outputting utterances from the clean corpus on one or more loudspeakers and recording the resulting sound field with the microphones . The baseline recognition system was trained on the clean training set, and achieved a word error rate (WER) of 6.19% on the clean test set. There are eight different "channels"—lapel, center, loosely coupled DAS and baseline recognizer, loosely coupled DAS and VTS, closely coupled DAS-BM and VTS, and another three similar ones for robust GSC. In the following recognition experiments, MAP adaptation was performed on the clean speech model using the cross-validation set for each channel-scenario pair, and then the adapted models were used to recognize the corresponding test set.

In our loosely and closely coupled model-based noise compensation experiments, a set of $K = 128$ diagonal gaussians was trained on the clean training set as the clean speech model in the log-FBE domain. For the loosely coupled model-based noise compensation, the VTS model-based compensation technique in [18] and [19] was used. It is similar to the compensation procedure described in Section II. The distinction is that there is no time-varying reference noises and no equalization filters, and instead the noise model is estimated using 20 frames from the nonspeech segment (ten frames at the utterance beginning, ten frames at the utterance end). In order to get the optimal compensation effect in the ML sense, the EM algorithm is taken to reestimate the noise model before doing compensation.

In the following, we first give the experimental results for the baseline recognizer using mono-microphone in Section IV-A. Then, the recognition results for various loosely/closely coupled schemes using array processors are discussed for DAS/DAS-BM and GSC, respectively, in Sections IV-B and IV-C. Finally, some discussions are presented in Section IV-D.

A. Experiments With Mono-Microphone

The first set of experiments was performed using mono-microphone (Lapel or Center) under different scenarios. The WER results are listed in Table I.

TABLE I
WER Results With Mono-Microphone (%)

Scenario	S1	S1S2	S1S2S3
Lapel	8.15	33.43	34.19
Center	10.43	66.18	83.51

TABLE II
WER Results With DAS/DAS-BM Array Processor (%)

Scenario	S1	S1S2	S1S2S3
Loosely coupled DAS and baseline recognizer	8.03	27.16	30.23
Loosely coupled DAS and VTS compensation	8.27	29.61	35.06
Closely coupled DAS-BM and VTS compensation	9.63	18.05	20.96

In this table, the "Lapel" row gives the WER results for the recordings from the desired speaker S1's lapel microphone (a close-talking setting), and the "Center" row gives the results for the recordings from the center table-top microphone (a distant-talking setting). Comparing these two rows of different microphone settings, we can see that as the distance from the target speaker increases, the center microphone is more susceptible to environmental interference such as the ambient noise and overlapping speech. For the three different scenarios, it is clear that the speech recognition performance becomes seriously degraded when there are several concurrent competing speakers. Even when the lapel microphone is placed very near the desired speaker, the problem still exists. So, these overlapping speech scenarios present great challenges to the baseline system with mono-microphone.

B. Experiments With DAS/DAS-BM Array Processor

In the second set of experiments, the DAS/DAS-BM array processor was used for spatial filtering. Speech recognition performances were compared for various loosely coupled and closed coupled schemes. The WER results are shown in Table II.

In spite of the simplicity of the DAS beamformer, the comparison of the "Loosely coupled DAS and baseline recognizer" row in Table II with the results in Table I shows its effectiveness for speech enhancement. For Scenario-S1, it can be seen that the table-top microphone array with DAS beamforming achieves comparable speech recognition performance with the close-talking lapel microphone. For the overlapping speech scenarios like S1S2 and S1S2S3, the competing speakers' speech is suppressed after DAS beamforming, and the recognition performances are improved over those using lapel microphone.

The "Loosely coupled DAS and VTS compensation" row corresponds to the experiments performed to investigate the combination of DAS array processing with model-based robust speech recognition in a loosely coupled approach. That is, the DAS beamformer operates in a MISO mode and its single enhanced speech output is fed directly into the speech recognizer with the VTS model-based noise compensation. Comparing with the above row, we can see that although the

TABLE III
WER RESULTS WITH ROBUST GSC ARRAY PROCESSOR (%)

Scenario	S1	S1S2	S1S2S3
Loosely coupled robust GSC and baseline recognizer	8.02	17.71	22.59
Loosely coupled robust GSC and VTS compensation	7.99	20.34	25.82
Closely coupled robust GSC and VTS compensation	8.28	15.17	18.39

speech recognizer is augmented with the VTS model-based noise compensation, this loosely coupled approach is not effective for improving the recognition performance. Note that the environmental interference is mainly nonstationary here, such as reverberation and overlapping speech from competing speakers. In this case, the time-invariant environment model cannot effectively represent this kind of nonstationary statistics, and thus the VTS model-based noise compensation fail to produce effective performance improvements.

The "Closely coupled DAS-BM and VTS compensation" row in Table II gives the recognition performances for close coupling DAS-BM and model-based noise compensation as shown in Fig. 4. In the experiments, the number of the BM outputs fed into the multichannel model-based compensation was set to 2, i.e., $M = 2$. For the overlapping speech scenarios like S1S2 and S1S2S3, this closely coupled scheme performs much better than the loosely coupled schemes (the "Loosely coupled DAS and baseline recognizer" row and the worse "Loosely coupled DAS and VTS compensation" row). These results are comparable with those reported in [10]. By exploiting the information provided by the reference noise outputs, the nonstationary residual noise is modeled more precisely, which in turn leads to more effective model-based noise compensation. For Scenario-S1, comparing with the "Loosely coupled DAS and baseline recognizer" row which uses the DAS beamformer output directly, we can see that after this multichannel model-based compensation, there is some performance degradation. In this scenario, there is no overlapping speech from competing speakers, and the target speech leaked through the BM becomes dominant in the reference channels. This violates the assumption made by (3) that the speech s_l and the reference noises, $\{r_l^j;\ j = 1,\ldots,M\}$, are mutually independent, and induces incorrect adaptation of the model parameters.

C. Experiments With Robust GSC Array Processor

The third set of experiments was performed to investigate the use of robust GSC as the front-end array processor. The WER results are summarized in Table III.

From the "Loosely coupled robust GSC and baseline recognizer" row in Table III, we can see that the overlapping speech from competing speakers is suppressed more effectively by robust GSC beamforming than by the simple DAS beamforming, and the recognition performances are further improved.

As shown in the "Loosely coupled robust GSC and VTS" row in Table III, loose coupling robust GSC array processing and VTS model-based noise compensation performs worse than directly feeding the GSC output to the baseline recognizer. This is similar to the above experimental results with DAS.

The "Closely coupled robust GSC and VTS compensation" row corresponds to the scheme shown in Fig. 6. For the overlapping speech scenarios like S1S2 and S1S2S3, the closely coupled scheme again achieves much better performance than the loosely coupled schemes. In addition, through the usage of MIAF processing (in combination with other robust adaptive beamforming techniques like AMC and NC), the target speech leakage is suppressed, and the performance for Scenario-S1 is guaranteed in this scheme. This result further shows the importance of appropriate combination of array signal processing techniques and multichannel model-based compensation techniques for improving recognition performance under various conditions.

D. Discussion

In the following, we give some examinations of the proposed closely coupled approach for better understanding of its behaviors. First, for our current use of the microphone array as a MIMO processor, we do not assume that one reference noise output corresponds to the estimate of one noise source in the sound field. We can block the desired speaker's speech using the simple Griffiths–Jim blocking matrix to generate the reference noise outputs, which then contain useful information about the environmental noise. Significant recognition performance improvements were obtained under the overlapping speech scenarios in the experiments. From this aspect, the closely coupled microphone array speech recognition system could perform without much concern about the precise one-to-one noise estimates.

Second, note that for the closely coupled approach, the leakage of the desired speech signal into the reference noise channels could mislead the subsequent model-based noise compensation. This problem is most serious under Scenario-S1.[6] In this case, it is less reasonable to assume that the speech and the reference noises are independent. However, when there are other competing speakers (e.g., S1S2 and S1S2S3 scenarios), the noise signals are dominant in the reference channels. In these cases, it is beneficial to exploit the reference noise outputs for model-based compensation, as shown by the performance improvements in the experiments. In addition, by use of robust GSC array processor, this problem can be alleviated to some extent. The recognition performance is guaranteed under Scenario-S1, through proper techniques (MIAF, AMC, and NC) to suppress target speech leakage.[7] To use more advanced MIMO algorithms for noise source segregation and estimation is worth further research.

[6] The target speech leakage in the BM outputs was measured to be -7 dB under Scenario-S1, averaging across the entire test utterances. For each utterance, the leakage value was calculated as the ratio in decibels between the average power of the BM outputs and that of the original microphone inputs.

[7] The target speech leakage in the MIAF output was measured to be -11 dB under Scenario-S1, averaging across the entire test utterances. For each utterance, the leakage value was calculated as the ratio in decibels between the average power of the MIAF output and that of the original microphone inputs.

Finally, note that the equalization filters used in the proposed closely coupled approach is a set of "short-time" linear filters, which is estimated frame by frame. Thus, the issue of compensating for long-term reverberation is not addressed in this paper.

V. Conclusion

In this paper, a new approach to microphone array speech recognition is proposed, in which the operations of the array processor and the speech recognizer are closely coupled. The array processor, as a MIMO module, generates not only the enhanced target speech signal but also some additional outputs that are informative about the background noises in the working environment. A multichannel environment model and a model-based noise compensation algorithm using VTS are proposed to make use of this multichannel information provided by the MIMO array processor. With the clean speech model and the environmental statistics, the compensation parameters are automatically estimated in the ML sense. Experimental results show that the proposed closely coupled approach can achieve environment modeling and acoustic model-based compensation more effectively than the conventional loosely coupled approach, especially to cope with the nonstationary interference under the overlapping speech situations.

Acknowledgment

The authors would like to thank Prof. Z. Wang for his support of this work, and M. Chen for conducting part of the experiments. They would also like to thank the reviewers and Prof. M. Ostendorf for their insightful comments which improved the manuscript.

References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. Berlin, Germany: Springer, 2001.

[2] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP–30, no. 1, pp. 27–34, Jan. 1982.

[4] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1375, Oct. 1987.

[5] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamforming for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

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

[7] S. Fischer and K. U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Commun.*, vol. 20, no. 3–4, pp. 215–227, Dec. 1996.

[8] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. ICASSP*, 1988, pp. 2578–2581.

[9] 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.

[10] D. Moore and I. A. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *Proc. ICASSP*, 2003, pp. 497–500.

[11] E. Lleida, J. Fernandez, and E. Masgrau, "Robust continuous speech recognition system based on a microphone array," in *Proc. ICASSP*, 1998, pp. 241–244.

[12] M. Omologo, M. Matassoni, and P. Svaizer, "Speech recognition with microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001, ch. 15, pp. 331–353.

[13] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 489–498, Sep. 2004.

[14] D. Raub, J. McDonough, and M. Wolfel, "A cepstral domain maximum likelihood beamformer for speech recognition," in *Proc. ICSLP*, 2004, pp. 817–820.

[15] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[16] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000, pp. 869–872.

[17] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Commun.*, vol. 24, no. 1, pp. 39–49, 1998.

[18] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environmental-independent speech recognition," in *Proc. ICASSP*, 1996, pp. 733–736.

[19] J. C. Segura, A. Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition: Experiments using the AURORA II database and tasks," in *Proc. EUROSPEECH*, 2001, pp. 221–224.

[20] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. EUROSPEECH*, 2001, vol. 2, pp. 1359–1362.

[21] N. S. Kim, "Time varying noise compensation using multiple Kalman filters," in *Proc. ICASSP*, 1999, pp. 1540–1543.

[22] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," in *Proc. ICASSP*, 2004, pp. 965–968.

[23] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. EUROSPEECH*, 2001, pp. 901–904.

[24] X. Zhao, Z. Ou, M. Chen, and Z. Wang, "Closely coupled array processing and model-based compensation for microphone array speech recognition," in *Proc. ICASSP*, 2005, pp. 417–420.

[25] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, ser. B, vol. 39, no. 1, pp. 1–38, 1977.

[26] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[27] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

Xianyu Zhao received the B.S. and M.S degrees in electronic engineering from the Harbin Institute of Technology, Harbin, Heilongjiang, China, in 1997 and 1999, respectively, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2005.

In 2005, he joined the France Telecom Research and Development Center, Beijing, China, and was involved in research work on speech signal processing. His research interests include speech and speaker recognition, intelligent signal processing, and machine learning.

Zhijian Ou (M'04) received the B.S. degree with the highest honor in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1998, and the M.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2000 and 2003, respectively.

Since 2003, he has been with the Department of Electronic Engineering, Tsinghua University, Beijing, China, as an Assistant Processor. His current research interests include signal processing, machine learning, with applications to speech and speaker recognition, and audio information retrieval.