COUPLING BEAMFORMING WITH SPATIAL AND SPECTRAL FEATURE BASED SPECTRAL ENHANCEMENT AND ITS APPLICATION TO MEETING RECOGNITION

Tomohiro Nakatani Mehrez Souden Shoko Araki Takuya Yoshioka Takaaki Hori Atsunori Ogawa

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237 Japan

ABSTRACT

This paper discusses microphone array based interference reduction approaches for robust automatic speech recognition. *A model based multichannel spectral enhancement* approach has recently been proposed for effectively reducing interference by exploiting both the spatial and spectral features of the signals. With the goal of further improving the effectiveness of this approach, we propose *a new framework* that combines this approach with *a microphone-array based beamforming approach*. Because the two approaches can work in a complementary manner in the proposed framework, they can greatly improve the interference reduction performance. We apply the proposed framework to the recognition of actual meetings, and show that it is superior to the use of beamforming or spectral enhancement alone in terms of the word error rates.

Index Terms: Speech enhancement, microphone array, model based approach, meeting recognition

1. INTRODUCTION

When we capture speech using distant microphones, various types of interference are mixed with the captured signals, thereby severely degrading automatic speech recognition (ASR) performance.

Microphone-array based beamforming has been studied as an approach that can reduce such interference in the captured signals [1, 2, 3, 4]. It controls the directivity pattern of the microphone array based on multichannel linear filtering so that it can extract signal components that come from the target talker's direction. Blind source separation (BSS) [5, 6] can be viewed as a technique that controls directivity patterns in a blind processing manner [7]. While beamforming can significantly reduce interference from point sources, the performance degrades when the interference arrives from many directions, due to the presence of diffuse noise and reverberation.

On the other hand, a model based *multichannel* spectral enhancement approach has recently been proposed for estimating clean speech spectra from a captured signal by exploiting the spectral and spatial features of the signals and their statistical models [8]. This approach is referred to as *DOminance based Locational and Power-spectral cHaracteristics INtegration (DOLPHIN)*. DOLPHIN is an extension of a factorial model based spectral enhancement approach [9, 10, 11], and can improve spectral enhancement accuracy by jointly utilizing the two features. For example, thanks to the use of the two features, spectral enhancement can be effectively achieved even when one of the features is not very reliable for the enhancement. One limitation of this approach is that the spectral enhancement does not modify the phase of the signals, and thus is not as effective as beamforming at canceling out point source interference.

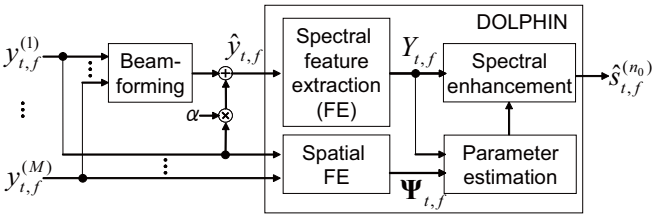

Fig. 1. Basic framework: Beamforming first reduces interference in the captured signal based on multichannel linear filtering. DOLPHIN then extracts the spectral feature, $Y_{t,f}$, from the beamforming output, and the spatial feature, $\mathbf{\Psi}_{t,f}$, from the multichannel observation. Then, DOLPHIN estimates the parameters of the source signal components included in the spectral feature, and finally applies spectral enhancement to the spectral feature to obtain the estimated speech, $\hat{s}_{t,f}^{(n_0)}$.

In this paper, to achieve better interference reduction using microphone arrays, we propose a new framework, in which both beamforming and DOLPHIN are utilized in a coupled system (see Fig. 1). In the framework, the beamforming first estimates a talker's speech based on captured signals by multichannel linear filtering, while DOLPHIN refines the speech spectral estimate of the utterance based on the spectral and spatial features. This framework allows the two methods to work in a complementary manner in that the point source interference can be well handled by the beamforming and the residual interference, including non-point source interference, can be effectively handled by DOLPHIN based on the two features. We show the superiority of the proposed framework by applying it to a meeting speech recognition task [12]. Note that in this framework, DOLPHIN needs to adapt to the amplitude modifications of the source components that the beamforming inevitably introduces. Thus, this paper also shows how DOLPHIN can adapt the channel parameters of the spectral model in an unsupervised manner based on an example configuration[1] of DOLPHIN.

It is important to note that researchers have been studying other frameworks for coupling beamforming with a model based *single-channel* spectral enhancement approach [13, 14]. These approaches can also provide better ASR performance than without the coupling. However, the spatial information obtained from the multichannel observation is not used directly to improve the spectral enhancement, except for the information obtained from the output of the beamformer. As a result, the performance of these approaches depends largely on the accuracy of the beamforming.

[1] The main contribution of this paper is the proposal of the framework. In the framework, we can adopt various configurations for DOLPHIN, which have been presented, for example, in [8].

2. COUPLING OF BEAMFORMING AND DOLPHIN

Suppose that N (≥ 1) talkers' utterances are captured by M (≥ 2) microphones jointly with ambient noise. Let t and f be time and frequency indices at a TF bin in the short time Fourier transform (STFT) domain, m ($= 1..M$) be a microphone index, and n ($= 0..N$) be a source index, where $n \geq 1$ represents one of N talkers and $n = 0$ represents the ambient noise. Then, the captured signal at a TF bin is modeled by

$$y_{t,f}^{(m)} = \sum_{n=0}^{N} x_{t,f}^{(n,m)}, \quad (1)$$

$$x_{t,f}^{(n,m)} = h_f^{(n,m)} s_{t,f}^{(n)} \quad \text{for} \quad n \geq 1, \quad (2)$$

where $y_{t,f}^{(m)}$ and $x_{t,f}^{(0,m)}$ are the STFTs of the captured signal and the ambient noise at the m-th microphone, $s_{t,f}^{(n)}$ and $x_{t,f}^{(n,m)}$ for $n \geq 1$ are the STFTs of the n-th talker's speech and its m-th microphone image, and $h^{(n,m)}$ is the acoustic transfer function from the n-th talker to the m-th microphone. Then, letting the 1-st microphone be *the reference microphone*, the goal of the speech enhancement for each talker n_0 is to obtain an estimate of $s_{t,f}^{(n_0)}$, denoted by $\hat{s}_{t,f}^{(n_0)}$, included in $y_{t,f}^{(1)}$ as *the n_0-th target signal*, where $x_{t,f}^{(n,1)}$ for $n \neq n_0$ and $x_{t,f}^{(0,1)}$ are regarded as interference signals to be reduced.

2.1. Basic framework: Single source extraction

To highlight the fundamental feature of the proposed framework, we start by discussing a simple version of the framework, referred to as the *basic framework*, which estimates utterances spoken by only one of the talkers, indexed by n_0. Later, we generalize it to an *advanced framework*, which can simultaneously handle multiple talkers.

Fig. 1 shows the processing flow of the basic framework. As indicated in the figure, the basic framework first conducts beamforming to estimate the n_0-th talker's utterances. The beamforming applies multichannel linear filtering to the multichannel observation as $\hat{y}_{t,f} = \sum_m w_f^{(n_0,m)} y_{t,f}^{(m)}$, where $w_f^{(n_0,m)}$ is a filter coefficient that enhances the n_0-th talker. The n_0-th talker's speech estimated based on beamforming can be represented by

$$\hat{y}_{t,f} = \sum_{n=0}^{N} \hat{x}_{t,f}^{(n)}, \quad (3)$$

$$\hat{x}_{t,f}^{(n)} = \hat{h}_f^{(n)} s_{t,f}^{(n)} \quad \text{for} \quad n \geq 1, \quad (4)$$

where $\hat{h}_f^{(n)} = \sum_m w_f^{(n_0,m)} h_f^{(n,m)}$ is a filtered acoustic transfer function, and $\hat{x}_{t,f}^{(0)} = \sum_m w_f^{(n_0,m)} x_{t,f}^{(0,m)}$ is the ambient noise remaining in $\hat{y}_{t,f}$. The above equations indicate that the beamforming only modifies the acoustic transfer functions of each source. The signal-to-interference ratio can be reduced by appropriately controlling the filter coefficients, $w_f^{(n_0,m)}$. For example, when $\hat{h}_f^{(n)} = 0$ is satisfied, the corresponding interference source is cancelled out. For such control, many techniques have been proposed for the blind estimation of filter coefficients [5, 6].

In the proposed approach, the output of the beamforming, $\hat{y}_{t,f}$, is then input into DOLPHIN as shown in Fig. 1, and treated as a reference signal, from which the target signal is estimated. By comparison with the case without beamforming, where DOLPHIN uses $y_{t,f}^{(1)}$ as the reference signal instead of $\hat{y}_{t,f}$, the spectral estimation can be more precise because $\hat{y}_{t,f}$ contains less interference than $y_{t,f}$. In addition, with this framework, we do not need to modify the processing of DOLPHIN by simply considering $\hat{y}_{t,f}$ as a microphone observation affected by different acoustic transfer functions.

Note that our preliminary experiments confirmed that it is also important to add the original reference signal multiplied with a certain weight α ($= 0.2$), namely $\alpha y_{t,f}^{(1)}$, to the output of the beamforming before it is input into DOLPHIN, as in Fig. 1. This is because it can avoid the case where certain target signal components are excessively reduced by the beamforming due to certain estimation errors.

2.2. DOLPHIN in basic framework

In this subsection, we present an example configuration of DOLPHIN to show how it can handle the signals in the basic framework. For this discussion, we adopt the simplest configuration of DOLPHIN for conciseness, where the spectral models are defined as Gaussian mixture models (GMM) for source log-spectra and no spatial models are utilized. The use of spatial models is discussed in Section 2.2.3 and the way of using other spectral models, such as GMMs for source mel-frequency cepstral coefficients (MFCC), can be found in [8, 15].

In Fig. 1, a spectral feature extraction (FE) block first extracts the spectral features, $Y_{t,f}$, from the reference signal, $\hat{y}_{t,f}$. Here, we adopt log-spectra as the spectral feature, which is obtained as

$$Y_{t,f} = \log |\hat{y}_{t,f}|^2. \quad (5)$$

To analyze the above feature, DOLPHIN introduces a fundamental assumption, namely each TF bin of the reference signal is dominated by one of the sources, and the spectral feature at the TF bin is equal to that of the dominant source. Letting $d_{t,f}$ be the index of the dominant source at each TF bin, referred to as the *dominant source index (DSI)*, and letting $X_{t,f}^{(n)}$, $S_{t,f}^{(n)}$, and $H_f^{(n)}$ be log-spectra of $\hat{x}_{t,f}^{(n)}$, $s_{t,f}^{(n)}$, and $\hat{h}_f^{(n)}$ defined as in eq. (5), this assumption is represented as

$$d_{t,f} = \arg\max_n \{X_{t,f}^{(n)}\}, \quad (6)$$

$$Y_{t,f} = X_{t,f}^{(d_{t,f})}, \quad (7)$$

where $X_{t,f}^{(n)}$ can be further decomposed based on eq. (4) into the n-th talker's speech, $S_{t,f}^{(n)}$, and its channel response, $H_f^{(n)}$, as

$$X_{t,f}^{(n)} = S_{t,f}^{(n)} + H_f^{(n)} \quad \text{for} \quad n \geq 1. \quad (8)$$

Then, we define the spectral model for each talker's speech by using a GMM as

$$p(S_{t,f}^{(n)}) = \sum_i w_i p(S_{t,f}^{(n)} | i_t^{(n)} = i), \quad (9)$$

$$p(S_{t,f}^{(n)} | i_t^{(n)} = i) = \mathcal{N}(S_{t,f}^{(n)}; \mu_{i,f}, \sigma_{i,f}). \quad (10)$$

Here, i is the Gaussian index, and w_i, $\mu_{i,f}$, and $\sigma_{i,f}$ are model parameters representing the mixture weight, the mean, and the variance of the i-th component, respectively. The model parameters are trained in advance, and can be speaker dependent or independent. The spectral model of the ambient noise is, on the other hand, modeled in this paper by a single Gaussian as

$$p(X_{t,f}^{(0)}) = \mathcal{N}(X_{t,f}^{(0)}; \mu_f^{(0)}, \sigma_f^{(0)}). \quad (11)$$

We assume the model parameters, $\mu_f^{(0)}$ and $\sigma_f^{(0)}$, can be estimated from speech absent segments of the reference signal[2].

[2]Parameters, $\mu_f^{(0)}$ and $\sigma_f^{(0)}$, can also be estimated jointly with the other parameters in the course of the EM iterations as discussed in [8].

2.2.1. Parameter estimation with channel adaptation

In the basic framework, unknown parameters to be estimated are the Gaussian index, $i_{t,f}^{(n)}$, and the channel response, $H_f^{(n)}$, for each spectral model[3], which are denoted as $\theta = \{\mathbf{H}, \mathbf{i}\}$, where each bold face symbol indicates a set of all parameters associated with the symbol. DOLPHIN maximizes an optimization function defined as $p(\mathbf{Y}, \mathbf{i}; \theta)$ to estimate θ, and it is accomplished based on the expectation-maximization (EM) algorithm, assuming that the DSIs, $d_{t,f}$, are handled as hidden variables. According to the discussion in [8], the auxiliary function for the EM algorithm can be defined and rewritten as

$$Q(\theta|\hat{\theta}) = E_{|\hat{\theta}}\{\log p(\mathbf{Y}, \mathbf{d}, \mathbf{i}; \theta)\} = \sum_n \sum_t Q_t^{(n)}(\theta^{(n)}|\hat{\theta}), \quad (12)$$

$$Q_t^{(n)}(\theta^{(n)}|\hat{\theta}) = \sum_f \Big\{ D_{t,f}^{(n)} p(X_{t,f}^{(n)} = Y_{t,f}|i_{t,f}^{(n)}; \theta^{(n)}) + (1 - D_{t,f}^{(n)}) \int_{-\infty}^{Y_{t,f}} p(X_{t,f}^{(n)}|i_t^{(n)}; \theta^{(n)}) dX_{n,f}^{(n)} \Big\} + \log p(i_t^{(n)}), \quad (13)$$

where $\theta^{(n)}$ is a subset of θ composed only of parameters associated with the n-th source, and $D_{t,f}^{(n)} = p(d_{t,f} = n|Y_{t,f}, \hat{i}_t^{(n)}; \hat{\theta}^{(n)})$ is the posterior of a DSI. In eq. (13), $p(X_{t,f}^{(n)}|i_t^{(n)})$ for $n \geq 1$ and $D_{t,f}^{(n)}$ can be rewritten, using the spectral model and the channel response, as

$$p(X_{t,f}^{(n)}|i_t^{(n)}; \theta^{(n)}) = p(S_{t,f}^{(n)} = X_{t,f}^{(n)} - H_f^{(n)}|i_t^{(n)}), \quad (14)$$

$$D_{t,f}^{(n)} = \frac{p(Y_{t,f}, d_{t,f} = n|\hat{i}_t^{(n)}; \hat{\theta})}{\sum_{n'} p(Y_{t,f}, d_{t,f} = n'|\hat{i}_t^{(n)}; \hat{\theta})}, \quad (15)$$

$$p(Y_{t,f}, d_{t,f} = n|\hat{i}_t^{(n)}; \hat{\theta}) = p(X_{t,f}^{(n)} = Y_{t,f}|\hat{i}_t^{(n)}; \hat{\theta}) \times \prod_{n' \neq n} \int_{-\infty}^{Y_{t,f}} p(X_{t,f}^{(n')}|\hat{i}_t^{(n)}; \hat{\theta}^{(n)}) dX_{t,f}^{(n')}. \quad (16)$$

The processing flow for estimating θ and the DSI posterior $D_{t,f}^{(n)}$ based on the EM algorithm is summarized in Algorithm I.

2.2.2. Spectral enhancement

After the parameter estimation, DOLPHIN estimates $X_{t,f}^{(n_0)}$ and $S_{t,f}^{(n_0)}$ of the n_0-th talker based on a minimum mean square error (MMSE) estimation as

$$\hat{X}_{t,f}^{(n_0)} = D_{t,f}^{(n_0)} \hat{Y}_{t,f} + (1 - D_{t,f}^{(n_0)}) \frac{\int_{-\infty}^{\hat{Y}_{t,f}} X_{t,f}^{(n_0)} p(X_{t,f}^{(n_0)}|\hat{i}_t^{(n_0)}; \hat{\theta}) dX_{t,f}^{(n_0)}}{\int_{-\infty}^{\hat{Y}_{t,f}} p(X_{t,f}^{(n_0)}|\hat{i}_t^{(n_0)}; \hat{\theta}) dX_{t,f}^{(n_0)}}, \quad (17)$$

$$\hat{S}_{t,f}^{(n_0)} = \hat{X}_{t,f}^{(n_0)} - \hat{H}_f^{(n_0)}. \quad (18)$$

The enhanced speech waveform can then be calculated by using an inverse Fourier transform of $\exp(\hat{S}_{t,f}^{(n_0)}/2)$ with the phase of the reference signal followed by overlap-add synthesis.

2.2.3. Incorporation of spatial features

As discussed in [8], the spatial features can be incorporated into DOLPHIN in an integrated manner. The incorporation improves the estimation of the DSI posterior, $D_{t,f}^{(n)}$, and thus improves the parameter estimation and the spectral enhancement by eqs. (17), (19), and (20).

[3] $S_{t,f}^{(n)}$ and $X_{t,f}^{(0)}$ can be marginalized out from $p(\mathbf{Y}, \mathbf{i}; \theta)$.

Algorithm I: Parameter estimation by DOLPHIN

1. Initialize the DSI posterior $D_{t,f}^{(n)}$.
2. Iterate the following until convergence is obtained
   (a) (M-step-1) Update the optimal GMM index of each source for $n \geq 1$ at each time frame t as
   $$\hat{i}_t^{(n)} = \arg\max_i Q_t^{(n)}(i_t^{(n)}|\hat{\theta}) \quad (19)$$
   (b) (M-step-2) Update the channel response, $\hat{H}_f^{(n)}$, for each source $n \geq 1$ based on the Newton-Raphson method as
   $$\hat{H}_f^{(n)} = \hat{H}_f^{(n)} - \left(\sum_t \frac{\partial^2 Q^{(n)}(\theta^{(n)}|\hat{\theta})}{(\partial H_f^{(n)})^2} \right)^{-1} \sum_t \frac{\partial Q^{(n)}(\theta^{(n)}|\hat{\theta})}{\partial H_f^{(n)}} \quad (20)$$
   where $\partial^2 Q_t^{(n)}/(\partial H_f^{(n)})^2$ and $\partial Q_t^{(n)}/\partial H_f^{(n)}$ are a gradient vector and a Hessian matrix of $Q_t^{(n)}$, respectively.
   (c) (E-step) Update the DSI posterior, $D_{t,f}^{(n)}$, for each source n and at each TF bin by eq. (15) (or by eq. (22)).

This paper adopts the same scheme as that used in [8, 16, 17] for the treatment of the spatial features. The spatial feature is defined as $\mathbf{\Psi}_{t,f} = \mathbf{y}_{t,f}/||\mathbf{y}_{t,f}||$, where $\mathbf{y}_{t,f} = [y_{t,f}^{(1)}, \ldots, y_{t,f}^{(M)}]$ and $||\cdot||$ is the Euclidean norm of a vector. Because this feature includes information on the level and phase differences between microphones, it can be used as the spatial feature. Based on the sparseness assumption [18], the spatial feature of the reference signal at each TF bin is assumed to be equal to that of the dominant source. Accordingly, the pdf of the spatial feature of the reference signal can be modeled by a mixture model as

$$p(\mathbf{\Psi}_{t,f}) = \sum_n z_n p(\mathbf{\Psi}_{t,f}|n; \lambda_f^{(n)}), \quad (21)$$

where $p(\mathbf{\Psi}_{t,f}|n; \lambda_f^{(n)})$ is a pdf of the spatial feature of the n-th source at frequency f, and z_n is the mixture weight. We adopt the pdf proposed in [16] for this model. With this model, we can estimate the model parameter, $\lambda_f^{(n)}$, for all speech sources $n \geq 1$ in a blind processing manner from the multichannel observation. It is also possible to estimate the model parameter for the ambient noise, $\lambda_f^{(0)}$, using a multichannel observation during the speaker absent segments[4].

With the spatial feature, the same parameter set, θ, is estimated by DOLPHIN as one that maximizes the optimization function defined as $p(\mathbf{Y}, \mathbf{\Psi}, \mathbf{i}; \theta)$. The estimation can still be accomplished based on the EM algorithm, where we only need to modify the update equation of the DSI posterior in eq. (15) as

$$D_{t,f}^{(n)} = \frac{p(\mathbf{\Psi}_{t,f}|n; \lambda_f^{(n)}) p(Y_{t,f}, d_{t,f} = n|\hat{i}_t^{(n)}; \hat{\theta})}{\sum_{n'} p(\mathbf{\Psi}_{t,f}|n'; \lambda_f^{(n)}) p(Y_{t,f}, d_{t,f} = n'|\hat{i}_t^{(n)}; \hat{\theta})}, \quad (22)$$

Based only on this modification, we can incorporate the spatial features into DOLPHIN to improve the spectral enhancement.

2.3. Advanced framework: Multiple source extraction

Suppose a case where beamforming can generate two or more outputs simultaneously corresponding to different talkers in the cap-

[4] Speaker absent segments can also be estimated from the captured signal based on the spatial features as discussed in [15].

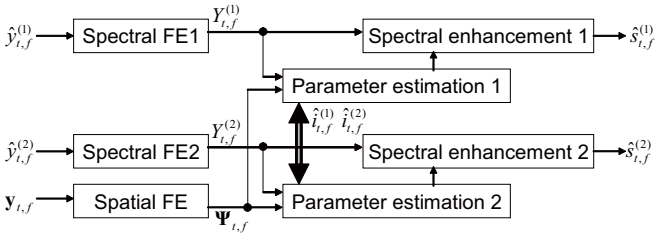

Fig. 2. Processing flow of DOLPHIN in advanced framework

Table 1. Specifications of meeting database [12]

	Office room	Soundproof room
T_{60} in ms	350 ms	120 ms
SNR	15 to 20 dB	20 to 25 dB
#Mics / #Talkers	8 / 4	
Mic.-talker distance	1 m	

tured signal. Many beamforming techniques are now available for this purpose [5, 6]. Even in this case, we can also adopt the basic framework to handle each output separately. However, this is obviously not optimal because each output contains the same set of speech sources, $s_{t,f}^{(n)}$, and it is more appropriate to use all the outputs for estimating each source at the same time. The advanced framework is introduced to achieve this.

Fig. 2 shows the processing flow of DOLPHIN for the advanced framework, where two reference signals, $\hat{y}_{t,f}^{(1)}$ and $\hat{y}_{t,f}^{(2)}$, are assumed to be input into DOLPHIN. With the advanced framework, the same optimization function as that of the basic framework is defined for each reference signal, and the optimization is conducted by maximizing the sum of all the functions, assuming that the Gaussian indices, $i_t^{(n)}$, of each source are shared by the functions. The same EM algorithm as in Algorithm I is derived for each reference signal except for the update of $\hat{i}_t^{(n)}$ in M-step-1. The updates of $\hat{i}_t^{(n)}$ can be obtained as those that maximize the sum of $Q_t^{(n)}$ over all the reference signals. Note that the same spatial features are shared by all the parameter estimation blocks as in Fig. 2.

3. EXPERIMENTS

We evaluated the proposed advanced framework using a meeting recognition task [12]. Table 1 summarizes the specifications of the meeting database used for the experiments. The meeting data were recorded in an office and a soundproof room, and consist of 68 sessions of meeting recordings (44 in the training set, 8 in the development set, and 16 in the test set). In each session, four talkers sit around a round table, and utterances generated by the talkers were recorded by 8 microphones located at the center of the table, referred to as table microphones. A headset was also used to record each talker's utterances for reference. Each session was about 15 minutes long, and each talker was assumed to be stationary at an unknown position during the session. The utterances generated by talkers may overlap, and are contaminated by stationary ambient noise, such as fan noise. The sampling frequency was 16 kHz. As indicated by Baseline in Table 2, the word error rates (WER) obtained by the SOLON recognizer [19] were very high when using the table microphones without any noise reduction preprocessing, and were much lower when we used the headsets as indicated by Headset in the table. This suggests that a major factor contributing to the high WERs of Baseline was the use of distant talking systems.

Table 2. WERs (%) of meeting recordings (test set) made in an office / a soundproof room.

	w/o AM adaptation	w/ AM adaptation
Headset	30.6 / 40.7	27.0 / 36.1
Baseline	86.5 / 79.8	80.5 / 79.0
ICA	60.6 / 59.6	49.5 / 48.1
MVDR	47.1 / 52.6	37.6 / 43.9
DOLPHIN	49.2 / 48.9	41.1 / 45.1
ICA+DOLPHIN	43.8 / **45.6**	37.9 / **41.6**
MVDR+DOLPHIN	**40.6** / 48.0	**35.5** / 42.7

For the beamforming, we adopted two different source separation methods, an independent component analysis (ICA) based BSS [5] and a minimum variance distortionless response beamforming (MVDR) based BSS [6]. For DOLPHIN, we adopted spectral models represented by GMMs of source MFCCs as proposed in [8, 15]. The window length and shift for DOLPHIN were set at 0.1 s and 0.025 s, respectively, and the dimensions of the filterbank and the MFCC were set at 40 and 13, respectively. We used the Corpus of Spontaneous Japanese (CSJ) [20] to train a talker independent spectral model, which was used for all the talkers. Because the CSJ was recorded with a headset, it has large channel mismatches with the meeting recordings. For the ambient noise, the spectral model was estimated from each meeting recording under evaluation in the course of the EM iteration as discussed in [8]. The spatial models were also estimated from each meeting recording under evaluation according to the methods described in Section 2.2.3.

For ASR, we used a speaker independent acoustic model trained on the CSJ based on a differential maximum mutual information (dMMI) criterion [21]. The language model was trained on transcriptions extracted from the CSJ, the training set of the meeting database, and web pages on the World Wide Web. The total vocabulary size of the ASR system was 156,000. The recognition was performed for each session with and without unsupervised acoustic model (AM) adaptation based on maximum-likelihood linear regression (MLLR) [22].

Table 2 summarizes the WERs obtained without any preprocessing (Baseline), with ICA, MVDR, or DOLPHIN, and with the advanced framework (ICA+DOLPHIN, MVDR+DOLPHIN). It clearly shows that each preprocessing method, namely ICA, MVDR, and DOLPHIN, greatly reduced the WERs, and the advanced framework further reduced the WERs. It is interesting to note that the two WERs obtained by the advanced framework are very close to each other although those obtained solely by beamforming are very different from each other. This suggests that DOLPHIN's use of the spatial and spectral features makes the performance of the proposed framework less dependent on that of the beamforming.

4. CONCLUDING REMARKS

This paper proposed an interference reduction framework using a microphone array for robust automatic speech recognition. The proposed framework, which is composed of beamforming and DOLPHIN, is very advantageous because the former reduces the interference by controlling the directivity patterns of microphone arrays while the latter reduces the residual interference by spectral enhancement based on both the spatial and spectral features of a multichannel observation. We experimentally showed that the WERs were greatly improved when we applied the proposed framework to a large vocabulary meeting recognition task with actual recordings.

5. REFERENCES

[1] J.L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 11, pp. 1508–1518, 1985.

[2] H. Cox, R.M. Zeskind, and T. Kooij, "Practical supergain," *IEEE Trans., Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 3, pp. 393–398, June, 1986.

[3] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *in Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[4] I.A. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," *in Proc. IEEE Int. Conf. Acost., Speech, Signal Process. (ICASSP'00)*, vol. 3, pp. 1723–1726, 2000.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[6] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for joint blind source separation and noise reduction," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'12)*, pp. 109-112, 2012.

[7] S. Araki, S. Makino, Y. Hinamono, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," *EURASIP Journal, Applied Signal Process.*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[8] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "Logmax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," *in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP'12)*, pp. 4029–4033. 2012.

[9] S.T. Roweis, "Factorial models and refiltering for speech separation and denoising," *in Proc. EUROSPEECH-2003*, pp. 1009-1012, 2003.

[10] S.J. Rennie, J.R. Hershey, and P.A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Magazine*, pp. 66–80, Nov. 2010.

[11] M.H. Radfar, W. Wong, R.M. Dansereau, and W.-Y. Chan, "Scaled factorial hidden Markov models: A new technique for compensating gain differences in model-based single channel speech separation," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'10)*, 2010.

[12] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 499–513, 2012.

[13] X. Zhao and Z. Ou, "Closely coupled array processing and model-based compensation for microphone array speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1114–1122, March, 2007.

[14] T. Yoshioka and T. Nakatani, "Time-varying residual noise feature model estimation for multi-microphone speech recognition," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'12)*, pp. 4913-4916, 2012.

[15] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, "Dominance based spectral and spatial feature integration for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.* (submitted), 2013.

[16] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[17] D.H. Tran-Vu and R. Häb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'10)*, 2010, pp. 241–244.

[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixture via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[19] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. Speech, Audio Process.*, vol. 15, no. 4, pp. 1352-1365, 2007.

[20] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," *in Proc. 2nd Int. Conf. Lang. Resources Evaluation (LREC'00)*, 2000, pp. 947–952.

[21] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," *in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP'10)*, 2010, pp. 4894–4897.

[22] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.