Feature Mapping of Multiple Beamformed Sources for Robust Overlapping Speech Recognition Using a Microphone Array

Weifeng Li, Longbiao Wang, Yicong Zhou, *Senior Member, IEEE*, John Dines, Mathew Magimai.-Doss, *Member, IEEE*, Hervé Bourlard, *Fellow, IEEE*, and Qingmin Liao

***Abstract*—This paper introduces a nonlinear vector-based feature mapping approach to extract robust features for automatic speech recognition (ASR) of overlapping speech using a microphone array. We explore different configurations and additional sources of information to improve the effectiveness of the feature mapping. First, we investigate the full-vector based mapping of different sources in a log mel-filterbank energy (log MFBE) domain, and demonstrate that retraining the acoustic model using the generated training data can help improve the recognition performance. Then we investigate the feature mapping between different domains. Finally in order to improve the qualities of the mapping inputs we propose a nonlinear mapping of the features from multiple beamformed sources, which are directed at the target and interfering speakers, respectively. We demonstrate the effectiveness of the proposed approach through extensive evaluations on the MONC corpus, which includes non-overlapping single speaker and overlapping multi-speaker conditions.**

***Index Terms*—Beamforming, microphone array, neural network, speech recognition, speech separation.**

I. Introduction

SPEECH overlap occurs frequently in natural conversations. For example, in a study on overlap in telephone conversations and multiparty meetings, it was found that 30-50% of all speech spurts include one or more frames of simultaneous speech by another talker [1]; In another study of 26 different meetings from the NIST meeting speech recognition evaluations, 12% of all foreground speaking time was overlapped by speech from one or more speakers [2]. Because of the detrimental effects of overlap, automatic recognition of speech in the presence of multiple simultaneous speakers - the so-called 'cocktail party' condition - remains a challenging problem (e.g. [1] [3], etc.). In such circumstances, headset microphones positioned next to the speakers' mouths have, to date, provided the best recognition performance, however they have a number of disadvantages in terms of cost and ease of use. The alternative is to capture the speech from one or more distant microphones located in the far field, however, such "remote microphone recordings" generally result in significantly reduced ASR performance.

Recent research has focused on techniques to efficiently integrate inputs from multiple distant microphones with the goal of improving the ASR performance. The most fundamental and important multi-channel method is the microphone array beamformer method, which consists of enhancing signals emanating from a particular location by combining the individual microphone signals. The simplest technique is the *delay-and-sum* (DS) beamformer, which compensates for delays to microphone inputs so that the target signal from a particular direction synchronizes, while noises from different directions do not. Other more sophisticated beamforming methods, such as the superdirective beamformer [4] and Generalized Sidelobe Canceller (GSC) [5], optimize the beamformer to produce a spatial pattern with a dominant response for the location of interest. The main limitation of these schemes is the issue of signal cancellation, which is more serious in the presence of overlapping speech.

It is important to note that the motivation behind microphone array techniques such as beamforming described above is to enhance or separate the speech signals, and as such they are not designed directly in the context of ASR. Particularly during periods of speaker overlap, improving the signal-to-noise ratio (SNR) of the signals captured through distant microphones may not necessarily be the best means of extracting features for robust ASR on distant microphone data [2], in which the target and interfering speech signals are mixed. This provides ample motivation for the investigation of other distant microphone processing techniques that specifically target the improvements of ASR performance.

Manuscript received October 09, 2013; revised April 28, 2014; accepted October 04, 2014. Date of publication October 20, 2014; date of current version November 14, 2014. This work was supported in part by Shenzhen Basic Research Grant JCYJ20120831165730913, in part by the Tateisi Science and Technology Foundation, in part by the Macau Science and Technology Development Fund under Grant 106/2013/A3, and in part by the Research Committee at University of Macau under Grants MYRG2014-00003-FST, MYRG113(Y1-L3)-FST12-ZYC, and MRG001/ZYC/2013/FST, and in part by European Union 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-033812) and the Swiss National Science Foundation through the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vincent Vanhoucke.

W. Li and Q. Liao are with the Shenzhen Key Laboratory of Information Science and Technology/Shenzhen Engineering Laboratory of IS&DRM, Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: li.weifeng@sz.tsinghua.edu.cn; liaoqm@sz.tsinghua.edu.cn).

L. Wang is with the Nagaoka University of Technology, Nagaoka 940-2188, Japan (e-mail: wang@vos.nagaokaut.ac.jp).

Y. Zhou is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yicongzhou@umac.mo).

J. Dines, M. M. Doss, and Hervé Bourlard are with the Idiap Research Institute, 1920-Martigny, Switzerland (e-mail: john.dines@idiap.ch; mathew@idiap.ch; herve.bourlard@idiap.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2364130

In this paper we introduce a novel feature mapping approach from multiple-microphone inputs specifically for the recognition of overlapping speech. In our feature mapping frameworks, the features of clean target speech are estimated (or approximated) by using the multiple distant speech sources. This is implemented implicitly and simultaneously by the following two phases: 1) mapping multiple mixed sources into one target source; and 2) mapping the distant sources into a clean target source. These two phases are explored by employing different configurations of additional sources of information in the context of automatic speech recognition of overlapping speech based on a microphone array. More specifically, we firstly introduce full-vector based mapping in the log mel-filterbank energy (log MFBE) domain. Since the speech energies of different speakers may lie in different filterbanks, it is possible to separate the target and interfering speeches in overlapping speech scenarios. Integrating the statistical distribution of clean speech with the properties of the de-correlations of Mel-frequency ceptral coefficients (MFCC), we then propose the mapping of the exacted features between different domains. Finally we propose a nonlinear mapping of features from the target and interfering distant sound sources to the clean target features, which leads to a non-linear processing (fusion) of the target speech and interfering speech. In this configuration two or three designed beamformers are directed at the target and interfering speakers, and a frequency domain binary mask post-filter is followed for obtaining the target and interfering speech more accurately. Experiments on the multi-channel numbers corpus (MONC) [6] show that our method yields a significant improvement in the ASR performance in overlapping speech scenarios, and can even avoid the adaptation steps which are commonly used in multi-condition ASR systems. We also demonstrate that the better quality of the estimated target and interfering speech as the inputs are helpful when using our non-linear feature mapping approach.

The paper is organized as follows. In Section II, we briefly describe related work. In Section III we present our proposed neural network based mapping approach and theoretically prove that minimizing the mean squared error (MMSE) of the static feature vectors also results in MMSE in their delta (acceleration) coefficients. In Section IV, we describe the experimental setup. From Section V to Section VII, we present the experimental studies using different mapping configurations, on a full-vector-wise mapping in the same log mel-filterbank energy (log MFBE) domain, on mapping from log MFBEs to MFCCs, and on the mapping from multiple sound sources. In Section IX, we summarize our main conclusions.

II. Related Work

In [7] a superdirective beamformer and further post-filtering were proposed to suppress interfering speech. However, in the case of overlapping speech (with coherent noise), the diffuse noise model used in the superdirective beamformers is inaccurate and may consequently introduce artifacts into the reconstructed signal. In [8] Kumatani *et al.* proposed an adaptive beamforming approach with a minimum mutual information criterion technique for the separation of overlapping speech. In their beamforming framework, one sub-band-domain beamformer in the GSC configuration was constructed for each source, and the active weight vectors of both GSCs were then jointly optimized to obtain two output signals with *minimum mutual information* (MMI), which is widely employed in blind source separation (BSS) [9] and independent component analysis (ICA) [10]. However these methods are basically linear methods, and to some degree their performance depends on the specified probability density functions (pdfs) of the Fast Fourier Transform (FFT) components of clean speech.

In [11], a *likelihood of maximizing beamforming* (LIMABEAM) was proposed to generate a sequence of features rather than a waveform. In this work a filter-and-sum beamformer structure is adopted, the beamformer is optimized by maximizing the likelihood of the correct hypothesis which comes from the speech recognition system. Their studies are directly applicable in the context of improving the performance of their automatic speech systems. Although some ASR improvement was shown on the condition of additive noise and reverberation, the algorithms result in a linear feature mapping approach which cannot recover the clean features very well.

While the beamforming methods result in a linear transformation, neural network (NN) [12] based mappings lead to a non-linear solution, and feature based mapping using neural networks has received considerable interest for robust ASR [13]–[17]. The idea of the feature mapping method is to obtain 'enhanced' or 'clean' features from the 'noisy' features extracted from the distant microphone recordings. The studies in [13][14] concentrated solely on the mapping of the original distant features to clean features. In [15]–[17], a microphone array is used and non-linear feature mapping of a DS enhanced speech signal to a clean speech signal is performed in the mel-frequency cepstral coefficient (MFCC) domain. In their mapping framework, a multi-layer perceptron (MLP) was trained for each MFCC component. We distinguish our approach by exploiting redundant or irrelevant information in a full-vector based mapping, using additional sources of information to improve the effectiveness of the mapping.

Recently deep learning based speech recognition [18]–[20] has received great interest. From the perspective of feature learning, the ideas and essences of our method and deep neural nets (DNN) for converting multiple noisy speech features to clean speech features are the same. Therefore the feature mapping studies in this paper can be viewed to be among deep learning based feature learning frameworks.

This paper is based on and extends our previous works [21]–[23]. In this paper we systematically explain the concept of our non-linear feature mapping approach and reformulate mathematically extracting multiple noisy feature vectors into one clean feature vector. More precisely, we modified the diagrams inside several figures to illustrate the different mapping frameworks more elaborately; we also perform the investigations and comparisons of the statistical distribution of log MFBEs and MFCCs of the clean speech, the generated training data, and the estimated test data; and several experiments are also added (e.g., vector-based mapping using array sources and DS beamformer source; using linear transform based mapping

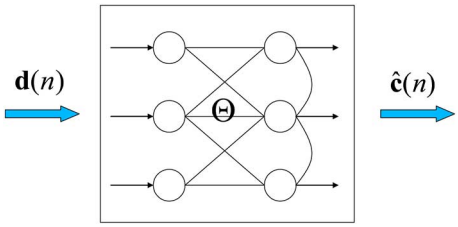


Fig. 1. Schematic representation of a network parametrized by Θ that should transform observations $\mathbf{d}(n)$ into estimated clean speech $\hat{\mathbf{c}}(n)$.

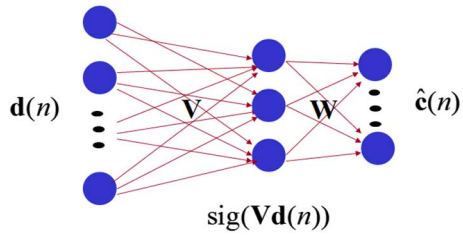


Fig. 2. A multilayer perceptron (MLP) network with one hidden layer, where $\mathbf{V}$ and $\mathbf{W}$ are weight matrices of input-layer and output-layer, respectively.

when using the feature vectors extracted from the center microphone speech and a DS beamformer, the exploration of the improvement of ASR performance in more serious overlapping speech scenario (with two interfering speakers) by using three beamformers, which are directed at the target speaker and the two interfering speech, etc.).

III. Feature Mapping Approach

Assume that we are given samples of feature vectors extracted from M 'noisy' distant microphone recordings at frame n, denoted by column-vectors: $\mathbf{d}_1(n), \mathbf{d}_2(n), \cdots, \mathbf{d}_M(n)$. If we concatenate them into a longer vector we have

$$\mathbf{d}(n) = [\mathbf{d}_1^T(n), \mathbf{d}_2^T(n), \cdots, \mathbf{d}_M^T(n)]^T. \tag{1}$$

Furthermore we consider a process that approximate the feature vector of clean speech $\mathbf{c}(n)$. In our mapping approach we take input features extracted from 'noisy' distant microphone recordings (either directly or after microphone array beamforming) and map these to 'clean' recordings (i.e., estimate the features of clean speech). This process may be implemented by a linear or non-linear transformation with the parameter set Θ (see Fig. 1). In the linear case, the feature vector of clean speech $\mathbf{c}(n)$ can be obtained by

$$\hat{\mathbf{c}}(n) = \mathbf{W}\mathbf{d}(n), \tag{2}$$

where the parameter set $\Theta = \{\mathbf{W}\}$ is obtained by minimizing the mean squared error:

$$\begin{aligned}\mathcal{E} &= \frac{1}{L}\sum_{l=1}^{L} ||\mathbf{c}(l) - \hat{\mathbf{c}}(l)||^2, \\ &= \frac{1}{L}\sum_{l=1}^{L} ||\mathbf{c}(l) - \mathbf{W}\mathbf{d}(n)||^2 \end{aligned} \tag{3}$$

over the training examples. Here, L denotes the number of training examples (frames). In terms of matrix notation, Eq. (3) can be written as [24][25]

$$\mathcal{E} = \frac{1}{L}\|\mathbf{C} - \mathbf{W}\mathbf{D}\|^2, \tag{4}$$

where $\mathbf{C} = [\mathbf{c}(1), \mathbf{c}(2), \cdots, \mathbf{c}(L)]^T$ and $\mathbf{D} = [\mathbf{d}(1), \mathbf{d}(2), \cdots, \mathbf{d}(L)]^T$ consist of training examples. The optimal $\mathbf{W}$ can be solved as

$$\hat{\mathbf{W}} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{C}. \tag{5}$$

In the non-linear case, we employ a multilayer perceptron (MLP) [12] with one hidden layer for implementing non-linear mapping. Formally, at n-th frame the i-th component of the feature vector of clean speech $\mathbf{c}(n)$ can be estimated using the MLP:

$$\hat{\mathbf{c}}(n) = \mathbf{W}sig((\mathbf{V}\mathbf{d}(n))\,, \tag{6}$$

where $\mathbf{V}$ and $\mathbf{W}$ are weight matrices of the input-layer and output-layer, respectively. $\mathrm{sig}(\cdot)$ is the sigmoidal activation function and has the form:

$$\mathrm{sig}(a) = \frac{1}{1+e^{-a}}. \tag{7}$$

Fig. 2 shows such a multiple-output multilayer perceptron (MLP) network. By minimizing Eq. (3) the optimal parameters $\Theta = \{\mathbf{V}, \mathbf{W}\}$ can be obtained through the error back-propagation algorithm [12], [26].

Note that clean speech is required for finding the optimal parameters in the neural network training, while in the test phase clean speech is no longer required, i.e., it is predicted from the input feature vectors from the enhanced target speech and the interfering speech.

With the assumption that the distribution of the target data is Gaussian-distributed, minimizing the mean square error in Eq. (3) is the result of the principle of maximum likelihood [26]. From the perspective of blind source separation (BSS) and independent component analysis (ICA), the principle of maximum likelihood, which is highly related to the minimization of mutual information between clean sources, can also be employed for estimating the clean sources [27]. Their methods, however, lead to a linear transformation, and the probability densities of the sources must be estimated correctly, while our mapping method can be highly non-linear and does not require the information concerning the probability densities of the sources.

On the other hand, delta and acceleration feature vectors have been proved to be efficient in improving ASR performance [28][29], and thus they are usually used in the recognizer. The delta feature vector (coefficients) at frame n are computed using the neighbor feature vectors from $n-K$ to $n+K$ frames in the following regression formula [30]:

$$\Delta\mathbf{c}(n) = \frac{\sum_{k=1}^{K} k[\mathbf{c}(n+k) - \mathbf{c}(n-k)]}{2\sum_k k^2}, \tag{8}$$

where $\mathbf{c}(n+k)$ and $\mathbf{c}(n-K)$ denote the corresponding static feature vectors at frame $(n+k)$ and $(n-K)$, respectively. We next theoretically prove that minimizing the mean squared error (MMSE) of the static feature vectors also results in MMSE for

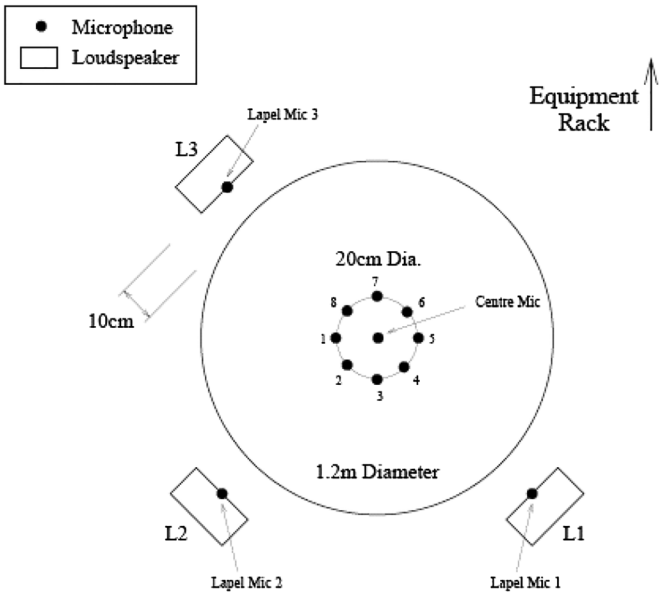


Fig. 3. The configuration of speech data recordings.

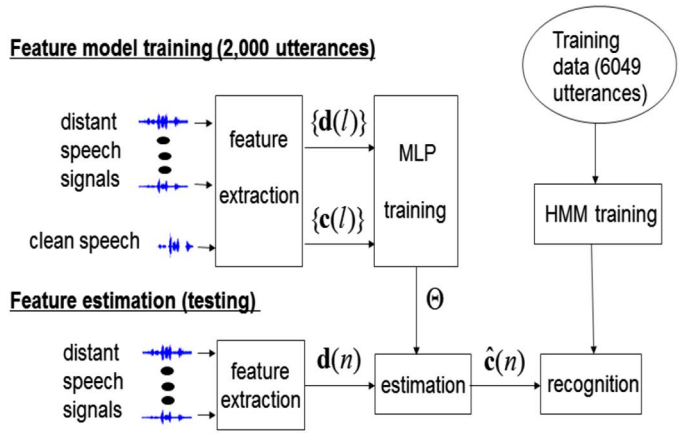


Fig. 4. Diagram of the feature mapping based speech recognition. $\{\mathbf{d}(l)\}$ and $\{\mathbf{c}(l)\}$ denote the training sets of the feature vectors extracted from distant microphone recordings and from clean speech, respectively. $\mathbf{d}(n)$ and $\hat{\mathbf{c}}(n)$ denote the feature vectors extracted from distant microphone recordings and the estimated feature vector of clean speech at n-th frame, respectively.

their delta coefficients (likewise for their acceleration coefficients), and thus we do not need to perform feature mapping for the delta and acceleration coefficients:

$$\begin{aligned}
&\min \sum\nolimits_n \|\Delta\mathbf{c}(n)-\Delta[\hat{\mathbf{c}}(n)]^2 \\
&=\min \frac{\sum_n \|\sum_k k\{[\mathbf{c}(n+k)-\hat{\mathbf{c}}(n+k)]+[\mathbf{c}(n-k)-\hat{\mathbf{c}}(n-k)]\}\|^2}{(2\sum_k k^2)^2} \\
&\approx \min \frac{\sum_n \sum_k k^2\{\|\mathbf{c}(n+k)-\hat{\mathbf{c}}(n+k)\|^2+\|\mathbf{c}(n-k)-\hat{\mathbf{c}}(n-k)\|^2\}}{(2\sum_k k^2)^2} \\
&\approx \min \frac{\sum_n 2\sum_k k^2\{\|\mathbf{c}(n+k)-\hat{\mathbf{c}}(n+k)\|^2\}}{(2\sum_k k^2)^2} \\
&=\min \frac{2\sum_k k^2 \cdot \sum_n\{\|\mathbf{c}(n+k)-\hat{\mathbf{c}}(n+k)\|^2\}}{(2\sum_k k^2)^2} \\
&=\min \frac{\mathcal{E}}{2\sum_k k^2},
\end{aligned}$$

where $\mathcal{E}$ denotes the cost function defined in Eq. (3). Here we assume the estimated errors, $[\mathbf{c}(n+k)-\hat{\mathbf{c}}(n+k)]$ and $[\mathbf{c}(n-K)-\hat{\mathbf{c}}(n-K)]$, are uncorrelated. Therefore our feature mapping of the static feature vectors leads to the minimization optimization of their delta and acceleration coefficients, which helps to improve the ASR performance.

IV. Experimental Data and Setup

The Multichannel Overlapping Numbers Corpus (MONC) [6] was used to perform speech recognition experiments. This database comprises a task for continuous digit recognition in the presence of overlapping speech. The configuration of speech data recordings in the MONC [6] is shown in Fig. 3. The database was collected in a moderately reverberant, $8.2\text{ m} \times 3.6\text{ m} \times 2.4\text{ m}$ rectangular room (reverberation time $RT_{60} \approx 0.5s$). Three loudspeakers (L1, L2, L3) were placed at 90 spacing around the circumference of a 1.2 m diameter circular table at an elevation of 35 cm. The placement of the loudspeakers simulated the presence of a desired speaker (L1) and two competing speakers (L2 and L3) in a realistic meeting room configuration. An 8-element, equally spaced, circular array of 20 cm diameter was placed in the middle of the table, and an additional microphone was placed at the center of the array. All subsequent discussion will refer to the recording scenarios as S1 (no overlapping speech), S12 (with 1 competing speaker L2), S13 (with 1 competing speaker L3), and S123 (with 2 competing speakers L2 and L3). The training data are equivalent to condition S1 of the development and evaluation sets.

The speech recognition experiments were carried out using whole-word HMMs. The word models had 16 emitting states, and each was modeled by a GMM of 20 components. The 'sil' and 'sp' models had three and one emitting state, respectively, with 36 Gaussian mixture components. The duration of the feature analysis was 25 msec with a frame shift of 10 msec. A 23-channel log-MFB analysis was applied and was transformed into 12 mel-frequency cepstral coefficients (MFCCs). Thus, the feature vector comprises 12 MFCCs and log-energy with corresponding delta and acceleration coefficients. HTK [30] was used for each feature extraction of the front-ends and training the acoustic model. In addition, maximum a posteriori (MAP) [31][32] adaptation was performed on these models using the development set for each scenario (thus, each adapted system comprised a set of four models, one adapted to each of the recording scenarios).

The corpus is divided into training data (6,049 utterances) and per-condition data sets for development/adaptation (2,026 utterances) and testing (2,061 utterances). In the feature mapping methods, the MLP is trained from data drawn from the development data set which consists of 2,000 utterances (500 utterances of each recording scenario in the development/adaptation set). The total number of training examples (frames) is 371,543. A diagram of the model training and feature estimation is given in Fig. 4. The size of the MLPs across the different ASR experiments were kept the same in this paper. Therefore the total number of parameters in the MLP was set up experimentally to be equal to 10% of the training frames [21].

V. Feature Vector Mapping in the Log Mel-Filterbank Energy Domain

We first estimated the log mel-filterbank energy (MFBE) vectors of clean speech by mapping those of distant speech[1]. The use of MMSE in the log spectral domain is motivated by the fact that the log spectral measure is more related to the subjective quality of speech [33] and that some better results have also been reported with log distortion measures [34][2]. On the other hand, in overlapping speech scenarios the speech energies of different speakers may lie in different filterbanks, which is advantageous to separate the target and interfering speeches through our feature learning methods. In [15]–[35], component-independent mapping is used, where an MLP corresponds to each component of a feature vector. Taking into consideration the log mel-filterbank energies are correlated, we propose to perform a full-vector based mapping via a universal MLP with the outputs comprising 23 log MFBE components and a log energy. We confirmed the improvement of speech recognition accuracy [21]. We performed two standard ASR experiments as our baselines:

1) center: Using the MFCCs extracted from the center microphone speech signal.
2) DS: Using the MFCCs extracted from the delay-and-sum (DS) beamformer enhanced speech signal.

Using the full-vector based mapping with MLP, we performed the following ASR experiments:

1) MA: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs extracted from all the 8-channel array speech as input, as shown in Fig. 4.
2) MDS: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs extracted from DS-enhanced speech as input.
3) MDSC: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs of both DS-enhanced speech and center microphone speech as inputs.

A diagram of the feature mapping based speech recognition using delay-and-sum beamformer and center speech with re-training the HMMs is shown in Fig. 5, in which Eq. (6) can be reformulated as

$$\begin{aligned}\hat{\mathbf{c}}(n) &= f(\mathbf{d}(n), \mathbf{s}(n)) \\ &= \mathbf{W}\mathrm{sig}\left(\mathbf{V}\begin{bmatrix}\mathbf{d}(n)\\ \mathbf{s}(n)\end{bmatrix}\right), \end{aligned} \tag{9}$$

where $\mathbf{d}(n)$ and $\mathbf{s}(n)$ denote the feature vectors extracted from the distant center microphone speech and the delay-and-sum beamformer at n-th frame, respectively. $\mathrm{sig}(\cdot)$ is the sigmoidal activation function. $\mathbf{V}$ and $\mathbf{W}$ are weight matrices of input-layer and output-layer, respectively. We also performed the mapping for the training data as well (and then re-training the acoustic model) to further reduce the mismatch between training and testing conditions. We refer to it as "+RT".

Table I shows the recognition results in terms of recognition accuracies. The upper and lower halves of this table depict

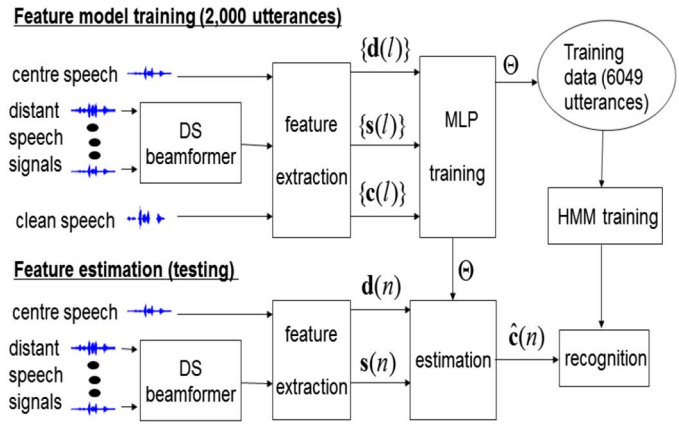


Fig. 5. Diagram of the feature mapping based speech recognition using delay-and-sum beamformer and center speech with re-training the HMMs. $\{\mathbf{d}(l)\}$, $\{\mathbf{s}(l)\}$, and $\{\mathbf{c}(l)\}$ denotes the training sets of the feature vectors extracted from center microphone speech, delay-and-sum beamformer, and clean speech, respectively. $\mathbf{d}(n)$ and $\mathbf{s}(n)$ denote the feature vectors extracted from center microphone speech, delay-and-sum beamformer at n-th frame, respectively. $\hat{\mathbf{c}}(n)$ denotes the estimated feature vector of clean speech at n-th frame.

TABLE I
Recognition Accuracies (as Percentages) of Different Systems for Vector-Based Mapping Studies. Upper Half of the Table Represents Accuracies for no Adaptation Case and Lower Half of the Table Represents Accuracies for Adaptation Case. The Best System Based Upon Average Accuracy Across all the Conditions is in Boldface Fonts

	S1	S12	S13	S123	Average
center	78.0	34.5	40.8	24.3	44.4
DS	73.8	46.3	54.7	39.8	53.7
MA	85.2	71.1	73.5	59.7	72.4
MDS	87.4	72.2	76.1	61.2	74.2
MDSC	88.0	76.1	79.4	64.8	77.1
MDSC+RT	88.6	78.9	83.8	72.5	**80.9**
MDSC+RT (linear)	86.5	73.2	79.0	66.7	76.4
center	89.0	38.7	46.9	27.6	50.6
DS	90.4	61.9	70.2	52.8	68.8
MA	85.2	72.4	76.7	61.9	74.1
MDS	89.1	72.8	77.1	63.0	75.6
MDSC	90.2	76.6	80.1	66.2	78.3
MDSC+RT	89.7	81.9	84.6	75.8	**83.0**
MDSC+RT (linear)	88.6	76.5	81.2	69.4	78.9

the recognition results without and with the adaptation of the acoustic models. Some of the major observations are:

- ASR performance drops when going from the single non overlap speaker condition S1 to the overlap speaker conditions S13, S12[3], and S123 with the three speaker overlap condition has the worst performance.
- "DS" is better than "center", and model level adaptation improves performance, which has also been observed previously in the literature [36].
- Irrespective of the method, the mapping approach always yields significant improvement in recognition accuracies for all conditions when compared with "center" and "DS" (except for the S1 condition after adaptation), with the improvements being significantly pronounced in the overlap conditions.

[1]The estimated log MFBEs are then transformed into MFCCs for recognition.

[2]In [34]. Porter and Boll found that for speech recognition, minimizing the mean squared errors in the log $|DFT|$ is superior to using all other DFT functions and to spectral magnitude subtraction.

[3]In the S12 condition the speakers are closer than the S13 condition which can explain why the S12 condition has a lower performance than the S13 condition.

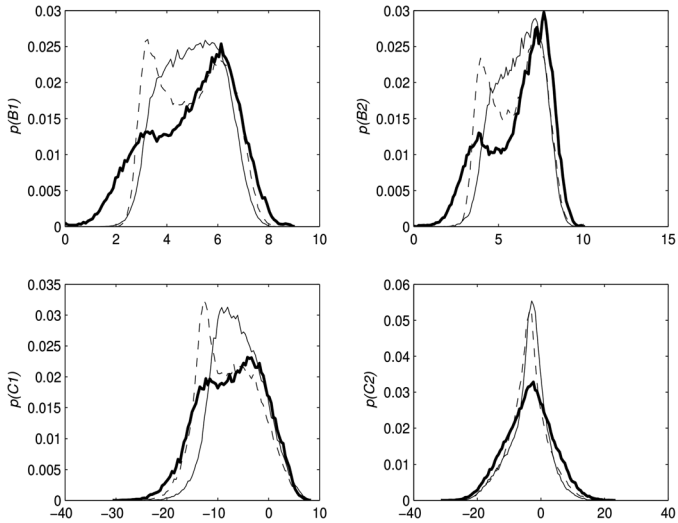


Fig. 6. Probability density functions (pdf) of the first and second order log MFBEs (upper half) and MFCCs (lower half) of the original clean training data (bold solid line), generated training data (dashed line), and estimated test data in S12 recording scenario (thin solid line).

- "MDS" performs better than "MA" suggesting that the quality of the features for mapping is important. "MDSC" performs better than "MDS" indicating that mapping the features and combining them from different "versions" of the speech signal at the input of the MLP is beneficial.
- For the MLP-based mapping methods, the feature adaptation for the training data and a subsequent re-training the acoustic model contributes to the improvement of the recognition performance in the overlapping speech scenarios. Fig. 6 shows the effect on the probability density functions (pdf) by adapting the training data of the acoustic model. It is observed that the mismatch of the probability density functions (pdf) between the training and test conditions is reduced by using the mapping-generated training data, rather than the original clean training data.
- Among the mapping methods "MDSC+RT" stands out as the best, demonstrating the effectiveness of incorporating combined features from different "versions" of speech signal as the input of the MLP, and re-training the acoustic model over the generated training data.

The last lines of the upper and lower halves of Table I indicate the recognition performance using linear mapping (i.e., Eq. (2)) rather than nonlinear mapping (i.e., Eq. (6)). It can be seen that non-linear mapping performs significantly better than linear mapping. Therefore non-linear mapping, and re-training the acoustic model over the generated training data, is adopted in the following studies.

VI. Feature Mapping Between Different Domains

In the above section, we estimated the log mel-filterbank energy (MFBE) vectors of clean speech by mapping those of distant speech. In the log MFBE domain, in overlapping speech scenarios the speech energies of different speakers may lie in different filterbanks, and the features are redundant and correlated, which are advantageous to separate the target and interfering speeches through our feature learning methods. In fact, the mapping need not to be performed between equivalent domains. From Fig. 6 (bold solid lines) it can be seen that the probability densities of the log MFBEs of the clean target are bi-modal (possibly because of the low SNR segments), rather than Gaussian. In this case, the maximum likelihood principle does not lead to minimizing the mean square error (MMSE), which we employed in Eq. (3). Therefore minimizing the mean square error (MMSE) in Eq. (3) will not be optimal if we perform the mapping in the log MFBE domain. Alternative mapping may be performed in the cepstral domain, where clean speech has an approximate Gaussian distribution (see Fig. 6), and the features are de-correlated and more straightforward in the context of speech recognition. It is advantageous to perform non-linear mapping of the features from the log mel-filterbank energy (log MFBE) domain to the features in the MFCC domain. We compared its ASR performance with those using the equivalent mapping in the MFCCs domain.

TABLE II
Recognition Accuracies (as Percentages) of the Mapping the Log MFBEs to MFCCs. Upper Half of the Table Represents Accuracies for No Adaptation Case and Lower Half of the Table Represents Accuracies for Adaptation Case

	S1	S12	S13	S123	Average
log MFBEs $\rightarrow$ MFCCs	89.5	81.0	84.5	74.0	**82.3**
MFCCs $\rightarrow$ MFCCs	88.2	77.5	82.6	71.2	79.9
log MFBEs $\rightarrow$ MFCCs	89.9	81.9	85.0	76.0	**83.2**
MFCCs $\rightarrow$ MFCCs	89.7	80.4	84.1	74.0	82.1

Table II shows the recognition results in terms of recognition accuracies compared to the equivalent mapping in the MFCCs domain. Here the best configuration in Table I, i.e. the non-linear mapping of the features vectors extracted from both the DS beamformer and the center microphone speech, is used. A re-training of the acoustic model over the generated training data is also included. The upper and lower halves of this table depict the recognition results without and with the adaptation of acoustic models. Some of the major observations are:

- The mappings from log MFBEs to MFCCs perform better than those from MFCCs, which confirms that the mixed filterbank inputs retain more information (e.g., different speech energy distributions over the filterbanks from the target and interfering speakers). Meanwhile, one can have the hypothesis that the smaller dynamic range of the log MFBEs as shown in Fig. 6 is advantageous for regression optimization [11].
- Additionally, compared with Table I, the mappings from log MFBEs to MFCCs perform better than those from log MFBEs to log MFBEs, which demonstrates that minimizing the mean square error (MMSE) in MFCCs domain is more advantageous than in the log MFBE domain. On the other hand, the properties of the de-correlations of MFCCs is helpful for speech recognition, but MFCCs are obtained by using a set of fixed discrete Cosine transforms (DCT). From Table I, it is suggested that our feature mapping methods can automatically de-correlate the log MFBE features, but in a more flexible way.
- The gains from model adaptation are marginal when we perform the feature mapping between different domains.

This may be explained by the fact that the mapping methods evaluated are already very effective at suppressing the influence of interfering speakers on the extracted features.

Therefore, the feature mapping from log MFBEs to MFCCs is employed in the following studies.

VII. Feature Mapping from Multiple Beamformed Sources

A. Proposed Method

In Section V, we found that augmenting the features to be mapped from the DS beamformer together with the center microphone could improve the mapping. The mapping method could be viewed as a non-linear processing technique that aims to approximate the clean speech through the fusion of the estimated (or enhanced) target speech and interfering speech[4]. If the qualities of the estimated target speech and interfering speech are improved, then it is highly possible that the clean speech can be approximated with greater precision. We pursued this idea further by mapping both the estimated target and interfering sound sources. The target and interfering audio signals are obtained by directing the DS beamformer in the directions of these sound sources. However, there still remain considerable undesired signal components in the DS outputs, thus, we further process them using a frequency-domain binary masking post-filter [37] to eliminate unwanted signal components. The frequency-domain masking post-filter is formulated as follows:

- If $b_i(k)$ is the frequency-domain output of the i-th beamformer for frequency bin k, the post-filtered output $p_i(k)$ is obtained as:

$$p_i(k) = h_i(k) b_i(k), \quad i = 1, \ldots, I \tag{10}$$

where the frequency response of the post-filter is estimated by:

$$h_i(k) = \begin{cases} 1 & \text{if } i = \arg\max_{i'} |b_{i'}(k)|, i' = 1, \ldots, I \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and I is the number of beamformers.

In S1[5], S12, S13 scenarios, the two beamformers (i.e., $I = 2$) are designed to correspond to the target and interfering speech. In the S123 scenario, the three beamformers (i.e., $I = 3$) are designed to be directed to the target speaker and the other two interfering speakers[6]. Therefore, in this section we propose to separate the target and interfering speech using DS beamforming followed by a frequency domain binary-masking based post-filter, and then perform our feature mapping method between different domains. In our experiments, once the beamformed speech signals were obtained, the 23-order log MFBEs were extracted and used as inputs in our mapping method to approximate the MFCCs of clean speech. A diagram of the feature mapping based speech recognition using two beamformers[7] with their subsequent poster-filtering is shown in Fig. 7, in which the r-th component of the MFCC feature vector of clean speech at frame n can be estimated by:

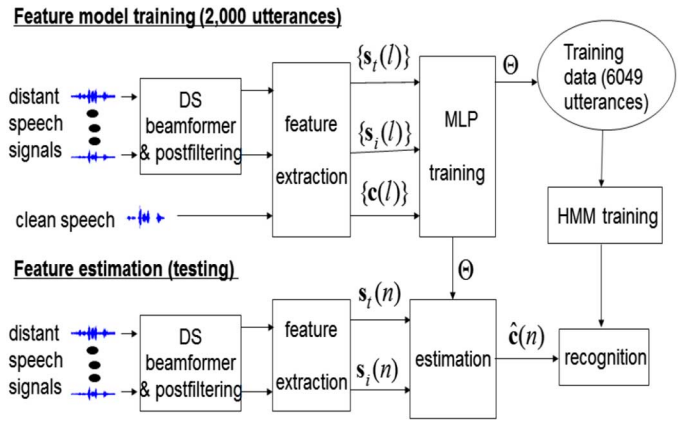


Fig. 7. Diagram of the feature mapping based speech recognition using two beamformers with their poster-filterings. Re-training the HMMs is also included. $\{\mathbf{s}_t(l)\}$ and $\{\mathbf{s}_i(l)\}$ denotes the training sets of the feature vectors extracted from the beamformers directed to target speaker and interfering speaker, respectively. $\{\mathbf{c}(l)\}$ denotes the training sets of the feature vectors extracted from clean speech. $\mathbf{s}_t(n)$ and $\mathbf{s}_i(n)$ denote the feature vectors extracted from the beamformers directed to target speaker and interfering speaker at n-th frame, respectively. $\hat{\mathbf{c}}(n)$ denotes the estimated feature vector of clean speech at n-th frame.

$$\begin{aligned} c_r(n) &= f(\mathbf{s}_t(n), \mathbf{s}_i(n)) \\ &= \sum_{p=1}^{P} \left(w_{p,r} \cdot sig\left(\mathbf{w}_{p,t}^T \mathbf{s}_t(n) + + \mathbf{w}_{p,i}^T \mathbf{s}_i(n) \right) \right), \end{aligned} \tag{12}$$

where $\mathbf{s}_t(n)$ and $\mathbf{s}_i(n)$ denote the log MFBE feature vectors extracted from the beamformers directed at the target and interfering speakers, respectively. $\mathbf{w}_{p,t}$ and $\mathbf{w}_{p,i}$ indicate their corresponding weights from the input layer to the p-th hidden neuron. $\mathbf{w}_{p,r}$ indicate the weights from the p-th hidden neuron to the r-th output.

B. Properties of the Proposed Method

We can now define the following properties of the proposed method.

1) From Eq. (12) the proposed non-linear feature mapping can be viewed as a generalized spectral subtraction in the feature domain[8], and the weights (or gains) are optimized using a minimum mean square error (MMSE) or maximum likelihood criterion [25][26].
2) The weights (or gains) are obtained by training the MLP universally on the collections of different overlapping scenarios (or number of sources), and in the test phase they can automatically adapt to different scenarios.

[4]The center microphone signal could be viewed as a mixture of the target and interfering noise.

[5]In the S1 scenario (only one active speaker), the secondary beamformer is directed to L3 as shown in Fig. 3, and thus the output of another beamformer is noise-like.

[6]Note that in this scenario the designation of the beamformers is different from our previous work [22], in which only two beamformers are used for the S123 scenario: one beamformer is directed at the target speech (L1 in Fig. 3) and the other directed at the middle position of the two interfering speakers (L2 and L3 in Fig. 3).

[7]three beamforms for the S123 scenario.

[8]i.e., the interfering components embedded in the beamformer directed to target speech may be subtracted by another beamformer directed to interfering speech.

TABLE III
RECOGNITION ACCURACIES (AS PERCENTAGES) OF DIFFERENT METHODS. UPPER HALF OF THE TABLE REPRESENTS ACCURACIES FOR NO ADAPTATION CASE AND LOWER HALF OF THE TABLE REPRESENTS ACCURACIES FOR ADAPTATION CASE. THE BEST SYSTEM BASED UPON AVERAGE ACCURACY ACROSS ALL THE CONDITIONS IS SHOWN IN BOLDFACE FONTS. THE RECOGNITION ACCURACIES SHOWN INSIDE THE PARENTHESIS DENOTE THE MAPPING FROM THE TWO BEAMFORMED SPEECH, WHICH IS USED IN [22]

	S1	S12	S13	S123	Average
DSmask	89.8	81.7	82.4	69.3	80.8
MmDS	90.7	86.9	88.3	84.2 (83.8)	87.6 (87.4)
MmDSmask	90.4	88.5	89.2	85.1 (84.6)	**88.3** (88.2)
DSmask	90.1	83.0	85.3	74.2	83.2
MmDS	90.9	87.5	88.8	85.0 (84.7)	88.0 (88.0)
MmDSmask	90.6	88.7	89.3	85.3 (85.1)	**88.5** (88.4)

3) The proposed method is established on the estimation of the target and interfering speech, which will be helpful in estimating the features of clean speech more accurately.
4) Generally speaking, we can universally employ two beamformers, directed at the target speaker and the other direction respectively. Unlike some blind source separation methods [9] (e.g., independent component analysis (ICA) [10]) the number of sources need not be prior knowledge.

C. Experimental Results

We performed the following ASR experiments:

1) DSmask: MFCCs extracted from the speech enhanced by the DS and subsequent masking post-filter.
2) MmDS: MFCCs estimated by the mapping of the two or three DS-enhanced speech sources;
3) MmDSmask: MFCCs estimated by the mapping of the two or three DS+masking enhanced speech sources.

Table III shows the recognition performance of the different experiments described above. In the S123 scenario, we also compared the ASR performance of the mapping from the three beamformed speech sources with that of the mapping from the two beamformed speech sources, shown in parenthesis in Table III. We can draw following inferences from the results:

- The frequency-domain masking post-filter is very effective at improving the quality of the separated speech (verified by informal listening). In the three speech-overlapping scenarios, the ASR performance is greatly improved by the frequency-domain masking post-filter.
- The mapping of the multiple DS-enhanced speech ("MmDS") sources yields significant improvement of the ASR performance compared with DS (especially without model adaptation), indicating that the interfering speech provides important information for mapping. Compared with Table II, the recognition accuracies are significantly improved. This suggests that estimating the interfering speech more accurately (using DS-enhanced interfering speech instead of the center microphone signal) is very helpful for the mapping method.
- The frequency-domain binary masking post-filter is also helpful for the mapping method. Except for the S1 condition, "MmDSmask" yields the best recognition system for overlap speech conditions. The well-estimated MFCC trajectories, as shown in Fig. 8, also illustrate the advantages of "MmDSmask". However, compared with "MmDS" the improvement of "MmDSmask" is not significant, which can be explained by the hypothesis that the MLP performs a similar role to the masking post-filter, both being provided with essentially the same information as the input (source and interfering speech) though in a different representation (FFT versus MFCC).
- In the S123 scenario, the mapping from the three beamformed speech sources (respectively directed to each of speakers) performs slightly better than that from the two beamformed speech sources. This demonstrates that the better qualities of the estimated interfering speech are very helpful for improving the ASR performance using the feature mapping method.
- Across the S1, S12, S13, and S123 scenarios, there is a significantly reduced mismatch between the four recordings. Moreover, for the mapping methods, the gains from model adaptation can be ignored. This may be explained by the fact that the feature mappings of multiple beamformed sources are very effective at approximating the target speech and suppressing the influence of interfering speakers on the extracted features. This avoids the need for adaptation to each scenario, which is required in the conventional multi-condition speech recognition systems.

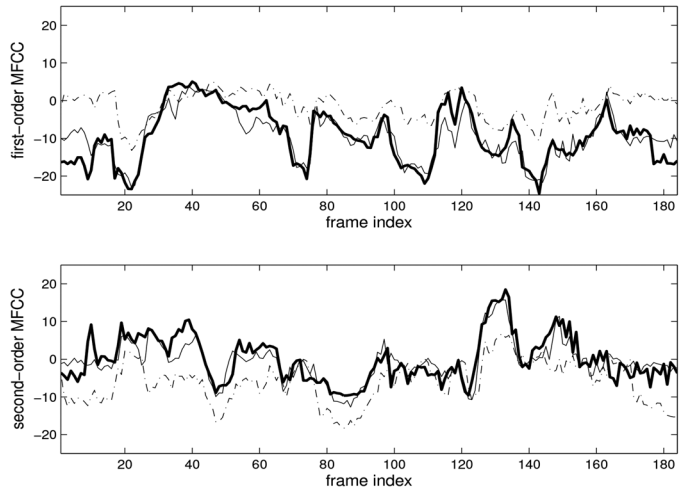


Fig. 8. Effect of the mapping method on the first and second MFCC trajectories in S12 recording scenario. bold solid line: MFCC trajectories of the clean speech; dash-dot line: MFCC trajectories of beamformed speech. thin solid line: the mapped MFCC trajectories.

VIII. DISCUSSIONS

Some issues concerning the proposed feature mapping method are worthy of investigating furthermore. In the following experiments, the adaptation parts are omitted for saving the space.

The proposed feature mapping method is based on MLP and is inherently non-linear. Some conventional feature transforms like feature-space maximum likelihood linear regression (fMLLR) [38] have been proven to be able to improve noise robustness of speech recognition. On the other hand, more recently deep neural networks (DNN) are employed to denoise the noisy speech [39][40]. For comparisons we performed the experiments using the supervised fMLLR and DNN. The same

TABLE IV
RECOGNITION ACCURACIES (AS PERCENTAGES) OF FEATURE-SPACE MAXIMUM LIKELIHOOD LINEAR REGRESSION (FMLLR) AND DENOISING AUTOENCODER (DAE). FOR COMPARISON "DSMASK" AND "MMDSMASK" ARE CITED FROM THE UPPER PART OF TABLE III. "MA" IS CITED FROM THE UPPER PART OF TABLE I

	S1	S12	S13	S123	Average
DSmask*	89.8	81.7	82.4	69.3	80.8
MmDSmask*	90.4	88.5	89.2	85.1 (84.6)	88.3
MmDSmask (linear)	86.2	83.4	84.7	72.3	81.7
DSmask+fMLLR	89.9	82.1	85.3	75.2	83.1
MA*	85.2	71.1	73.5	59.7	72.4
DAE	85.5	72.0	74.1	61.2	73.2

TABLE V
RECOGNITION ACCURACIES (AS PERCENTAGES) OF SUPER-DIRECTIVE BEAMFORMER WITH A SUBSEQUENT BINARY MASKING POST-FILTER (SDMASK) AND A COMBINATION OF SDMASK WITH FEATURE-SPACE MAXIMUM LIKELIHOOD LINEAR REGRESSION (FMLLR) AND OUR MLP FEATURE MAPPING (MMSDMASK)

	S1	S12	S13	S123	Average
SDmask	90.4	82.2	82.4	73.8	82.2
SDmask+fMLLR	90.5	84.1	86.3	79.4	85.1
MmSDmask	90.6	88.7	89.5	84.9	88.4

TABLE VI
RECOGNITION ACCURACIES (AS PERCENTAGES) OF TRAINING MLP ON THE S1 AND S12 CONDITIONS BUT TESTING ON THE S13 AND S123 CONDITIONS. "DSMASK" AND "MMDSMASK" ARE CITED FROM THE UPPER PART OF TABLE III FOR COMPARISON

	S1	S12	S13	S123	Average
DSmask*	89.8	81.7	82.4	69.3	80.8
MmDSmask*	90.4	88.5	89.2	85.1 (84.6)	88.3
MmDSmaskS	90.5	88.9	87.1	65.2	82.9
SDmaskS+fMLLR	90.6	84.8	85.0	61.2	80.4
MmSDmaskS	90.7	89.3	87.1	65.7	83.2

2,000 utterances from the development data set are used for training. In the fMLLR case the adaptation features are based on MFCCs obtained from "DSmask" in Section VII. In the DNN case a denoising autoencoder (DAE) with three hidden layers (500 neurons per layer) similar to [39] is trained over the 2000-utterance noisy-clean speech pairs[9]. After pretraining in a layer-by-layer manner, all the layers are stacked to form a denoising autoencoder for fine tuning. Note that our proposed method in Section VII are the non-linear mappings of two (or three) beamformed sources, while fMLLR applies a linear adaptation on a single beamformed source only. DAE is a denoising neural network with the inputs of eight original noisy sources. As shown in Table IV, fMLLR performs better than "DSmask" and "MmDSmask (linear)"[10]in the upper part of Table III, but not as well as the proposed "MmDSmask". Although compared with "MA" in the upper of Table I, DAE provides slightly better performance but with a higher computation cost, it performs significantly worse than "MmDSmask". These demonstrate the effectiveness of the proposed non-linear mapping of multiple beamformed sources.

Simple delay-and-sum (DS) beamformers are used in the experiments aforementioned. In [41][42], it is shown that a super-directive (SD) beamformer with a subsequent frequency-domain binary masking post-filter in Eqs. (10) and (11) can consistently yield super performance in meeting scenarios. We therefore incorporated it in our proposed method, denoted by "MmSDmask" in Table V. Like in Section VII two super-directive beamfomers directed to the target speech and interferers are used as the inputs of MLP. Table V shows the recognition performance with this setup. Indeed, super-directive beamformer with a subsequent binary masking post-filter yields better performance than delay-and-sum beamfomer with a same post-filter. A feature-space maximum likelihood linear regression (fMLLR) can also help improve the performance. However like "MmDSmask" vs "MmDS" in the upper part of Table III, the gains of "MmSDmask" from "MmDSmask" are quite marginal. This demonstrates that the proposed mapping method is not dependent on the qualities of multiple beamformed sources.

Generalization to unseen conditions is worth investigating for supervised learning algorithms, and thus we design the following experiments to testify MLP's generalization abilities. An MLP with two directed delay-and-sum (DS) beamformers followed by a binary masking post-filter is trained over 1,000 (500 each from S1 and S12 scenarios) utterances which are from the development data set. The test conditions consist of all the four scenarios (S1, S1S2, S1S3, and S123). As shown in Table VI, in average this setup (denoted by "MmDSmaskS") performs better than "DSmask". Compared to the original "MmDSmask", "MmDSmaskS" provides marginally better performance for S1 and S12 scenarios (matched conditions), which are used for training MLP. As for the unseen (or mismatched) S13 scenario, "MmDSmaskS" performs slightly worse than "MmDSmask" but performs significantly worse for the unseen (or mismatched) S123 scenario. These tendencies also occur when using the super-directive (SD) beamforming with a subsequent binary masking post-filter ("SDmaskS+fMLLR" and "MmSDmaskS" in Table VI). This may be explained because the inputs of MLP for the S13 scenario are similar to those of the S12 scenario while the inputs of MLP for the S123 scenario differ far from those for the S12 scenario, in which the trained MLP weights are not capable to recover the clean features.

The data used for the experiments aforementioned were recorded in the same room, and thus they have the same acoustic characteristics. In order to testify whether the proposed method can work in a different acoustic condition, 200 utterances are recorded using a circular microphone array (20 cm diameter) as same as MONC database [6] but in a $7.2\ \text{m} \times 5.0\ \text{m} \times 3\ \text{m}$ rectangular room (reverberation time $RT_{60} \approx 0.85s$) in Nagaoka University of Technology, Japan, as shown in Fig. 9. 200 different clean utterances are displayed through three loudspeakers in Fig. 9. A multi-channel recording device "Tokyo Electron device TD-BD-16ADUSB" was used for recording. In this experimental setup, MLP weights are trained on MONC database [6] while the test data are from this new environment (denoted by "MmDSmask2" in Table VII).

[9]The inputs of the DAE are the log MFBEs of eight noisy channels, and the outputs are the MFCCs of clean speech.

[10]i.e., linear mapping (using Eq. (2)) the features from multiple DS-beamformed sources (with a binary masking post-filter).

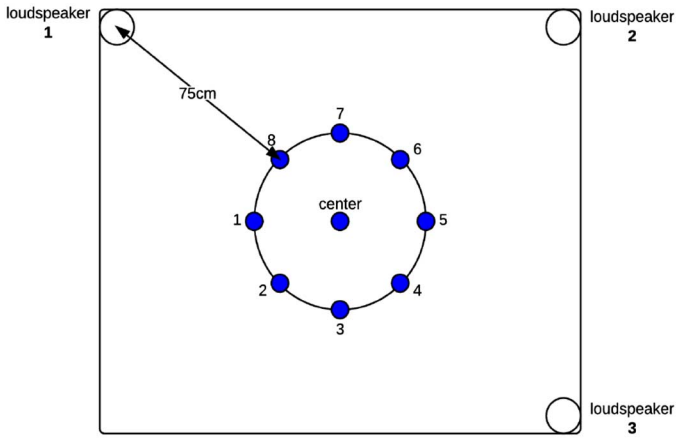


Fig. 9. The configuration of speech data recorded in Nagaoka University of Technology, Japan.

TABLE VII
RECOGNITION ACCURACIES (AS PERCENTAGES) OF NEWLY RECORDED DATA. FOR ALL THE METHODS, THEIR ACOUSTIC MODELS ARE TRAINED OVER MONC DATABASE [6]. FOR "SDMASK2+FMLLR2" AND "MMDSMASK2" THE MLP WEIGHTS ARE TRAINED ON MONC DATABASE [6]. FOR COMPARISON "DS" IS CITED FROM THE UPPER PART OF TABLE I. "DSMASK" AND "MMDSMASK" ARE CITED FROM THE UPPER PART OF TABLE III. "SDMASK+FMLLR" IS CITED FROM TABLE V. NOTE THAT FOR "DS", "DSMASK", "SDMASK+FMLLR", AND "MMDSMASK" THEIR TRAININGS AND TESTINGS ARE PERFORMED UNDER MATCHED CONDITIONS, WHILE "FOR DS2", "DSMASK2", "SDMASK2+FMLLR2", AND "MMDSMASK2" THEIR TRAININGS AND TESTINGS ARE PERFORMED UNDER MISMATCHED CONDITIONS

	S1	S12	S13	S123	Average
DS*	73.8	46.3	54.7	39.8	53.7
DSmask*	89.8	81.7	82.4	69.3	80.8
SDmask+fMLLR*	90.5	84.1	86.3	79.4	85.1
MmDSmask*	90.4	88.5	89.2	85.1 (84.6)	88.3
DS2	69.6	29.8	36.6	24.3	39.9
DSmask2	75.5	39.2	42.7	37.5	49.0
SDmask2+fMLLR2	75.8	42.5	45.9	42.4	51.7
MmDSmask2	72.4	50.3	51.2	47.1	55.3

"DS2" and "DSmask2" denote DS beamformer and with a binary masking post-filter respectively, which are applied to the newly recorded data with their acoustic models trained over MONC database [6]. The original "DS", "DSmask", "SDmask+fMLLR", and "MmDSmask" in MONC database [6] are cited for comparison. Note that for "DS", "DSmask", "SDmask+fMLLR", and "MmDSmask" their trainings and testings are performed under matched conditions, while "for DS2", "DSmask2", "SDmask2+fMLLR2", and "MmDSmask2" their trainings and testings are performed under mismatched conditions. As shown in Table VII, for all the methods the recognition results in this new acoustic environment are significantly worse than those in MONC database (under matched conditions). The reasons may be that the reverberation time in the new environment is remarkably longer than that in the MONC environment, and the speech qualities recorded by multi-channel recording device "Tokyo Electron device TD-BD-16ADUSB" are not as high as MONC data. "SDmask2+fMLLR2" and "MmDSmask2" still outperform "DS2" and "DSmask2", however its recognition accuracies are far lower than those of the original "SDmask+fMLLR" and "MmDSmask". This may be explained by the fact that the inputs of MLP in the new environments are different from those used for training MLP, and the originally trained MLP weights are no longer feasible in the new environment. To address this problem, we will investigate training an MLP over the speech recorded from various acoustic environments in the future.

IX. CONCLUSIONS AND FUTURE WORK

We have presented our approach to improving the recognition performance of overlapping speech using a non-linear feature mapping method. We first employed the full-vector based mapping in log mel-filterbank energy (log MFBE) domain and improved recognition accuracy by re-training the acoustic model over the generated training data. We then improved the recognition accuracy by exploring the mapping of the extracted features between different domains. Finally the best recognition performance was achieved by using a microphone array to extract the features from the directions of the target and interfering sound sources, which was followed by mapping these features to those of clean speech. The proposed approach achieved considerable improvements in ASR performance for overlapping multi-speaker conditions, and was also effective for the single non-overlapping condition. We discovered that our proposed approach resulted in a non-linear processing (fusion) of the target and interfering speech, and the improved qualities of the estimated target and interfering speech were very helpful in improving the ASR performance using our proposed non-linear feature mapping method. We also demonstrated that the well estimated feature vectors (i.e., MFCCs), obtained via our final proposed method, could avoid the need for adaptation to a particular recording scenario.

There are several areas where further investigation is needed. In the MONC corpus, the clean speech is available, however in real applications actual clean speech is not readily available, and instead close-talking microphones (CTM) are usually employed. It is worth investigation the recognition performance of the our proposed mapping method using CTM speech. We plan to extend this work to more realistic environments (e.g., overlapping speech encountered in meeting scenarios), and detect speaker overlap and non-overlap regions in multi-party meetings and train/adapt the MLP directly using close-talking microphone speech as target speech.

REFERENCES

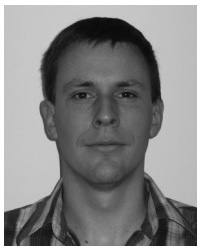
[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech*, 2001, pp. 1359–1362.
[2] O. Cetin and L. Shriberg, "Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap," in *Proc. ICASSP*, 2006, pp. 357–360.
[3] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 84–91, Jan. 2005.
[4] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[5] 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.
[6] "The multichannel overlapping numbers corpus," *The Hidden Markov Model Toolkit* [Online]. Available: http://www.idiap.ch/mccowan/arrays/monc.pdf
[7] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proc. ICASSP*, 2003, pp. 497–500.

[8] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. W. , "Adaptive beamforming with a minimum mutual information criterion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2527–2541, Nov. 2007.
[9] S. Haykin, *Unsupervised Adaptive Filtering*. New York, NY, USA: Wiley, 2000.
[10] T.-W. Lee, *Independent component analysis: Theory and applications*. Boston, MA, USA: Kluwer, 1998.
[11] M. L. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Dept. of Elect .and Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2003.
[12] S. Haykin, *Neural networks - a comprehensive foundation*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1998.
[13] H. Sorensen, "A cepstral noise reduction multi-layer neural network," in *Proc. ICASSP*, 1991, pp. 933–936.
[14] B. de Vries *et al.*, "Neural network speech enhancement for noise robust speech recognition," in *Proc. Int. Workshop Applicat. Neural Netw. Telecomm.*, 1995.
[15] C. Che, Q. Lin, J. Pearson, B. de Vries, and J. Flanagan, "Microphone arrays and neural networks for robust speech recognition," in *Proc. Workshop Human Lang. Technol.*, 1994, pp. 342–347.
[16] Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. de Vries, J. Pearson, and J. Flanagan, "Robust distant-talking speech recognition," in *Proc. ICASSP*, 1996, pp. 21–24.
[17] D. Yuk, C. Che, L. Jin, and Q. Lin, "Environment independent continuous speech recognition using neural net works and hidden Markov models," in *Proc. ICASSP*, 1996, pp. 3358–3361.
[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[19] L. Deng, J. Li, J. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 8604–8608.
[20] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, May 2013, pp. 7398–7402.
[21] W. Li, M. Magimai.-Doss, J. Dines, and H. Bourlard, "mlp-based log spectral energy mapping for robust overlapping speech recognition," in *Proc. 16th Eur. Signal Process. Conf. (EUSIPCO)*, 2008.
[22] W. Li, J. Dines, M. Magimai.-Doss, and H. Bourlard, "Neural network based regression for robust overlapping speech recognition using microphone arrays," in *Proc. Interspeech*, 2008, pp. 2012–2015.
[23] W. Li, J. Dines, M. Magimai.-Doss, and H. Bourlard, "Non-linear mapping for multi-channel speech separation and robust overlapping speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP '09)*, 2009, pp. 3921–3924.
[24] V. S. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*. New York, NY, USA: Wiley, 1998.
[25] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning*. : Springer New York Inc., 2001.
[26] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[27] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, 2000.
[28] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 2783–2793, Oct. 1994.
[29] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.
[30] S. Young *et al., The Hidden Markov Model Toolkit* [Online]. Available: http://htk.eng.cam.ac.uk/
[31] Y. Gotoh, M. M. Hochberg, and H. F. Silverman, "Using map estimated parameters to improve HMM speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. I, pp. 229–232.
[32] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
[33] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[34] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. ICASSP*, 1984, pp. 18.A.2.1–18.A.2.4.
[35] W. Li, C. Miyajima, T. Nishino, K. Itou, K. Takeda, and F. Itakura, "Adaptive nonlinear regression using multiple distributed microphones for in-car speech recognition," *IEICE Trans. Fundam. Electron., Commun., Comput. Sci.*, vol. E88-A, no. 7, pp. 1716–1723, 2005.
[36] A. Stolcke *et al.*, "The sri-icsi spring 2007meeting and lecture recognition system," in *Lecture Notes in Comput. Sci.*, 2007.
[37] I. McCowan, M. H. Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio-visual sensor array," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2005, pp. 1382–1385.
[38] g. Saon and J. Chien, "Large vocabulary continuous speech recognition systems: A look at some recent advances," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 18–33, Nov. 2012.
[39] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.
[40] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 3512–3516.
[41] E. Zwyssig, F. Faubel, S. Renals, and M. Lincol, "Recognition of overlapping speech using digital MEMS microphone arrays," in *Proc. ICASSP*, 2013, pp. 7068–7072.
[42] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proc. MLMI*, 2007, pp. 295–305.

Weifeng Li received the M.E. and Ph.D. degrees in information electronics at Nagoya University, Japan, in 2003 and 2006, respectively. He joined the Idiap Research Institute, Switzerland, in 2006, and in 2008 he moved to the Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland, as a Research Scientist. Since 2010, he has been an Associate Professor in the Department of Electronic Engineering/Graduate School at Shenzhen, Tsinghua University, China. His research interests lie in the areas of audio and visual signal processing, Biometrics, Human–Computer Interactions (HCI), and machine learning techniques. He is a member of the IEEE and IEICE.

Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. and Dr.Eng. degree from Toyohashi University of Technology, Japan, in 2005 and 2008, respectively. From July 2000 to August 2002, he worked at the China Construction Bank. He was an Assistant Professor in the faculty of Engineering at Shizuoka University, Japan, from April 2008 to September 2012. Since October 2013, he has been an Associate Professor at Nagaoka University of Technology, Japan. His research interests include robust speech recognition, speaker recognition and sound source localization. He received the "Chinese Government Award for Outstanding Self-financed Students Abroad" in 2008. He is a member of the IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Yicong Zhou (M'07–SM'14) received his B.S. degree from Hunan University, Changsha, China, and his M.S. and Ph.D. degrees from Tufts University, Massachusetts, USA, all degrees in electrical engineering. Dr. Zhou is currently an Assistant Professor in the Department of Computer and Information Science at University of Macau, Macau, China. His research interests focus on multimedia security, image/signal processing, pattern recognition and medical imaging. Dr. Zhou won the third prize of Macau Natural Science Award in 2014 and is a member of SPIE (International Society for Photo-Optical Instrumentations Engineers).



John Dines received the B.E. degree in electrical and electronic engineering from the University of Southern Queensland, Australia, in 1998 and received the Ph.D. degree from the Queensland University of Technology, Australia, in 2003 for the thesis "Model-based trainable speech synthesis and its applications." Between 2003 and 2013, he was with the Idiap Research Institute, Martigny, Switzerland, where he was working mostly in the domain of rich transcription. His other research interests include statistical parametric speech synthesis and leveraging progress in synthesis and recognition to make advances across domains. From 2010 to 2013, he was founding CTO of internet start-up Koemei SA, also based in Martigny. At present, he is an independent consultant in speech technology and is an external collaborator with the Speech and Hearing group at the University of Sheffield. He is the author of over 50 international journal and conference publications and three patents.

Mathew Magimai.-Doss (S'03–M'05) received the bachelor of engineering (B.E.) in instrumentation and control engineering from the University of Madras, India, in 1996, the master of science (M.S.) by research in computer science and engineering from the Indian Institute of Technology, Madras, India, in 1999, the PreDoctoral diploma and the docteur ès sciences (Ph.D.) from Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2000 and 2005, respectively. He was a postdoctoral fellow at International Computer Science Institute (ICSI), Berkeley, USA, from April 2006 till March 2007. Since April 2007, he has been a Research Scientist at the Idiap Research Institute, Martigny, Switzerland. His research interests include speech processing, automatic speech recognition, automatic speaker recognition, spoken language processing, signal processing, statistical pattern recognition and artificial neural networks. He is the author of over 75 peer-reviewed journal and conference publications. He is an associate editor of the IEEE SIGNAL PROCESSING LETTERS.

Hervé Bourlard (M'89–SM'95–F'00) received the electrical and computer science engineering degree and the Ph.D. degree in applied sciences from the Faculté Polytechnique de Mons, Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Martigny, Switzerland, Full Professor at the Swiss Federal Institute of Technology at Lausanne (EPFL), and Director of a National Center of Competence in Research in "Interactive Multimodal Information Management." Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of the ICSI Board of Trustees. His main interests are in signal processing, statistical pattern classification, multichannel processing, artificial neural networks, and applied mathematics, with applications to speech and natural language modeling, speech and speaker recognition, computer vision, and multimodal processing. He is the author/co-author/editor of four books and over 250 reviewed papers (including one IEEE paper award) and book chapters. He is an IEEE Fellow for "contributions in the fields of statistical speech recognition and neural networks." He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of ICASSP 2002, General Chairman of Interspeech 2003), and on the editorial board of several journals (e.g., past co-Editor-in-Chief of *Speech Communication*). Over the last 20 years, he has initiated and coordinated numerous large international research projects, as well as multiple collaborative projects with industries. He is an appointed expert for the European Commission and, from 2002 to 2007, was also part of the European Information Society Technology Advisory Group (ISTAG).

Qingmin Liao received the B.S. degree in radio technology from the University of Electronic Science and Technology of China, Chengdu, China, in 1984, and the M.S. and Ph.D. degrees in signal processing and telecommunications from the University of Rennes 1, Rennes, France, in 1990 and 1994, respectively. Since 1995, he has been joining with Tsinghua University, Beijing, China. In 2002, he became Professor in the Department of Electronic Engineering of Tsinghua University. Since 2010, he has been the director of the Division of Information Science and Technology in the Graduate School at Shenzhen, Tsinghua University. He is also affiliated with the Shenzhen Key Laboratory of Information Science and Technology (Director), China. Over the last 30 years, he has published over 100 peer-reviewed journal and conference papers. His research interests include image/video processing, transmission and analysis; biometrics; and their applications to teledetection, medicine, industry, and sports.