Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language 00 (2016) 1–18

www.elsevier.com/locate/csl

An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech

Yan-Hui Tu[a], Jun Du[a,*], Qing Wang[a], Xiao Bao[a], Li-Rong Dai[a], Chin-Hui Lee[b]

[a] *University of Science and Technology of China, Hefei, Anhui, PR China*
[b] *Georgia Institute of Technology, Atlanta, GA, USA*

Received 8 April 2016; received in revised form 3 December 2016; accepted 4 December 2016

Abstract

We present an information fusion approach to the robust recognition of multi-microphone speech. It is based on a deep learning framework with a large deep neural network (DNN) consisting of subnets designed from different perspectives. Multiple knowledge sources are then reasonably integrated via an early fusion of normalized noisy features with multiple beamforming techniques, enhanced speech features, speaker-related features, and other auxiliary features concatenated as the input to each subnet to compensate for imperfect front-end processing. Furthermore, a late fusion strategy is utilized to leverage the complementary natures of the different subnets by combining the outputs of all subnets to produce a single output set. Testing on the CHiME-3 task of recognizing microphone array speech, we demonstrate in our empirical study that the different information sources complement each other and that both early and late fusions provide significant performance gains, with an overall word error rate of 10.55% when combining 12 systems. Furthermore, by utilizing an improved technique for beamforming and a powerful recurrent neural network (RNN)-based language model for rescoring, a WER of 9.08% can be achieved for the best single DNN system with one-pass decoding among all of the systems submitted to the CHiME-3 challenge.

Keywords: CHiME challenge; Deep learning; Information fusion; Microphone array; Robust speech recognition

1. Introduction

With the emergence of eyes-busy and hands-free speech-enabled applications on multi-microphone portable devices, the robust recognition of microphone array speech in distant-talking scenarios has become one of the most critical issues to be addressed for the massive deployment of spoken language services. For the past several decades, many techniques (Gong, 1995; Li et al., 2014a) have been proposed to handle this challenging problem, but there have not been many performance benchmarks for studying noise robustness issues. One remarkable benchmark was the Aurora series initiated by Nokia in 2000, including the Aurora-2 (Pearce and Hirsch, 2000), Aurora-3 (Aurora,

* Corresponding author.
E-mail address: tuyanhui@mail.ustc.edu.cn (Y.-H. Tu), jundu@ustc.edu.cn (J. Du), xiaosong@mail.ustc.edu.cn (Q. Wang), baox@mail.ustc.edu.cn (X. Bao), lrdai@ustc.edu.cn (L.-R. Dai), chl@ece.gatech.edu (C.-H. Lee).

http://dx.doi.org/10.1016/j.csl.2016.12.004
0885-2308/ 2017 Published by Elsevier Ltd.

1999; 2000; 2001a; 2001b) and Aurora-4 (Hirsch, 2002) tasks. The Aurora-2 and Aurora-4 databases were designed with artificially generated noisy data for the recognition tasks of small and medium-sized vocabularies, respectively, whereas the Aurora-3 task aimed at recognizing digit strings in real automobile environments.

Evolving into the mobile era and with the ever-increasing popularity of the deep learning technologies, the focus on noise robustness has been reinvigorated by a recent series of CHiME challenges (Barker et al., 2013; 2015; Vincent et al.,2013) in recent years. This series differs from the Aurora tasks in several aspects. First, the scenarios are extended to far-field automatic speech recognition (ASR) in everyday listening environments, e.g., the family living room. Second, the room impulse responses (RIRs) simulating speaker movements and reverberation conditions have been convolved with the utterance to generate more realistic artificial noisy data. Third, research on microphone array based ASR has been more emphasized than conventional single-microphone techniques. One main difference in the CHiME-3 challenge, which was launched in 2015, from the previous CHiME-1 and CHiME-2 challenges was the use of a set of real-world data collected from several typical scenes via a mobile tablet device equipped with microphone arrays.

In this sense, the CHiME-3 challenge might serve a new research direction attempting to solve ASR problems in real-world applications. The initially released official results also indicated that conventional techniques that worked well on the simulation data could fail on real data. Among all of the systems submitted to CHiME-3, several categories of solutions were proposed. The multi-channel speech enhancement approaches based on beamforming techniques (Yoshioka et al., 2015; Hori et al., 2015; Sivasankaran et al., 2015; Zhao et al., 2015; Heymann et al., 2015; Jalalvand et al., 2015; Pang and Zhu, 2015; Prudnikov et al., 2015; Mousa et al., 2015; Barfuss et al., 2015) have been widely used as the mainstream. In Hori et al. (2015), Sivasankaran et al. (2015) and Prudnikov et al. (2015), the super-directive minimum variance distortionless response (MVDR) beamformer provided in the official baseline system (Barker et al., 2015) was replaced with a robust delay and sum beamformer for the real data. To improve the MVDR beamformer, the Top-1 system (Yoshioka et al., 2015) adopted a spectral mask-based approach to obtain accurate estimates of the acoustic beam-steering vectors, and (Zhao et al., 2015) proposed a cross-correlation and eigen-decomposition method for microphone gain estimation. For the post-filtering techniques of beamforming, spatial coherence filtering (Pang and Zhu, 2015; Barfuss et al., 2015) or filtering for dereverberation (Yoshioka et al., 2015; Mousa et al., 2015) was commonly used. In Jalalvand et al. (2015), several beamforming techniques were combined at the lattice level during decoding. Single-channel deep learning based front-end processing was investigated in Bagchi et al. (2015) and Ma et al. (2015). Bagchi et al. (2015) used a deep neural network (DNN)-based spectral mapping method that predicted clean filter bank features from noisy spectra, and Ma et al. (2015) conducted DNN-based mask estimation using pitch-based features. However, the algorithms could not yield significant performance improvements for the final systems. In Du et al. (2015), we proposed a solution via a large neural net consisting of subnets with different architectures, namely, deep neural networks (DNNs) (Vesel et al., 2013) and recurrent neural networks (RNNs) (Graves, 2012), to combine multiple knowledge sources by early feature fusion and late score fusion. Overall, the NTT system (Yoshioka et al., 2015) achieved the best performance on this challenging task, which indicates that an effective front end via conventional beamforming techniques incorporated with single-channel deep learning based approaches for acoustic modeling in the back-end is a successful solution for multi-channel speech recognition.

Our proposed framework consists of early and late fusion stages. In early fusion, diverse features are concatenated to compensate for imperfect beamforming. First, a concatenation of multi-channel acoustic features is investigated, with each channel corresponding to one beamforming result of a channel subset in a microphone array. This is quite different from the conventional approach in which one single overall output, after beamforming combines all of the channels of the array, is fed to the recognizer. One reason that such a proposed multi-channel feature concatenation technique can achieve a better performance might be that it reduces the risk caused by the imperfection of the existing beamforming approaches, especially for a microphone array with many highly diverse channels.

A few issues must be considered carefully in the proposed fusion approach. First, for multi-channel concatenation, there is an increase in the input layer size for the DNN, which can be even larger than the hidden layers and often leads to performance degradation. To alleviate this problem, multiple-frame expansion is applied to the main channel, whereas only one central frame is used for the other channels. Second, appending multiple enhanced features is believed to be beneficial, motivated by the observation that the use of the enhanced features from the main channel alone could not provide an improvement over the noisy features on the real data, possibly due to the large residual noise (Barker et al., 2015). Different feature normalization approaches, speaker-related features, and auxiliary features are also studied in early fusion. For late fusion, the outputs of all subnets with different architectures are combined via a simple posterior average strategy (Li and Sim, 2013) to generate a single output set for subsequent decoding. Based on our experiments, the early and late fusions are equally important and strongly complementary in terms of reducing the ASR word error rates (WERs). The proposed two-stage fusion may be superior to either pure early fusion or late fusion. If all the information is concatenated in early fusion, then it is difficult to handle the issue of high dimensionality in the input layer and dynamic ranges of different features. Similarly, if only late fusion is used, the poor performance of each subnet can be predictable.

The main contributions of this study, on top of our previous work submitted to CHiME-3 (Du et al., 2015) are (i) a simplified version of the MVDR approach from Yoshioka et al. (2015) with a comparable recognition performance is adopted as the most effective one among all of those used in beamforming; (ii) in terms of boosting the input signal-to-noise ratio (SNR), the beamforming techniques are not necessarily the same as those used in training and recognition stages in our new framework, which can relax the constraint that the same front-end techniques should be applied to training and testing in the conventional pattern recognition framework; and (iii) more detailed descriptions of the implementations, new experiments, and an expanded discussion of results are provided. Finally, by using an improved MVDR approach and language model re-scoring, we can achieve the best recognition performance for any DNN system with one-pass decoding among all of the submissions to the CHiME-3 challenge.

The remainder of the paper is organized as follows. Section 2 describes the CHiME-3 challenge task. In Section 3, we give a detailed system description of our proposed deep learning framework. Early fusion and late fusion are elaborated in Sections 5 and 6, respectively. In Section 7, we report experimental results, and we conclude our findings in Section 8.

2. The CHiME-3 challenge task

The CHiME-3 challenge is designed to focus on real-world and commercially motivated scenarios in which a person is talking to a mobile tablet device in a variety of real and challengingly public conditions (Barker et al., 2015). Four environments have been selected: a café (CAF), a street junction (STR), a public transport (BUS) and a pedestrian area (PED). For each environment, two types of noisy speech data have been provided: real and simulated. The real data are collected from 6-channel recordings of speakers reading the same sentences from the WSJ0 corpus (Garofalo et al., 2007) in four environments. The simulated data are constructed by mixing clean utterances with environmental noise recordings by using the techniques described in Vincent et al. (2007). For the ASR evaluation, the data are divided into official training, development and test sets.

The development and test data consist of 410 and 330 utterances, respectively, with the same text as the corresponding sets in the WSJ0 5k task. Each sentence is read by four different talkers in one randomly selected environment, for totals of 1640 (410 × 4) and 1320 (330 × 4) real development and test utterances. Similarly, simulated data are generated for the development and test sets. The training data include 1600 real noisy utterances from the

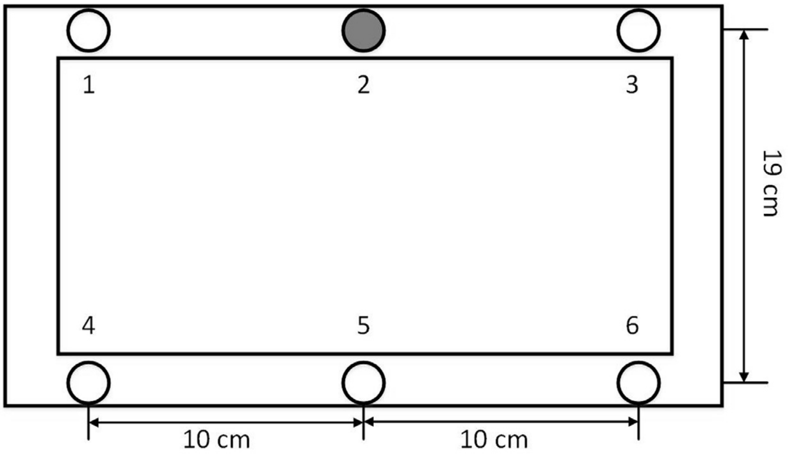


Fig. 1. Geometry of microphone array.

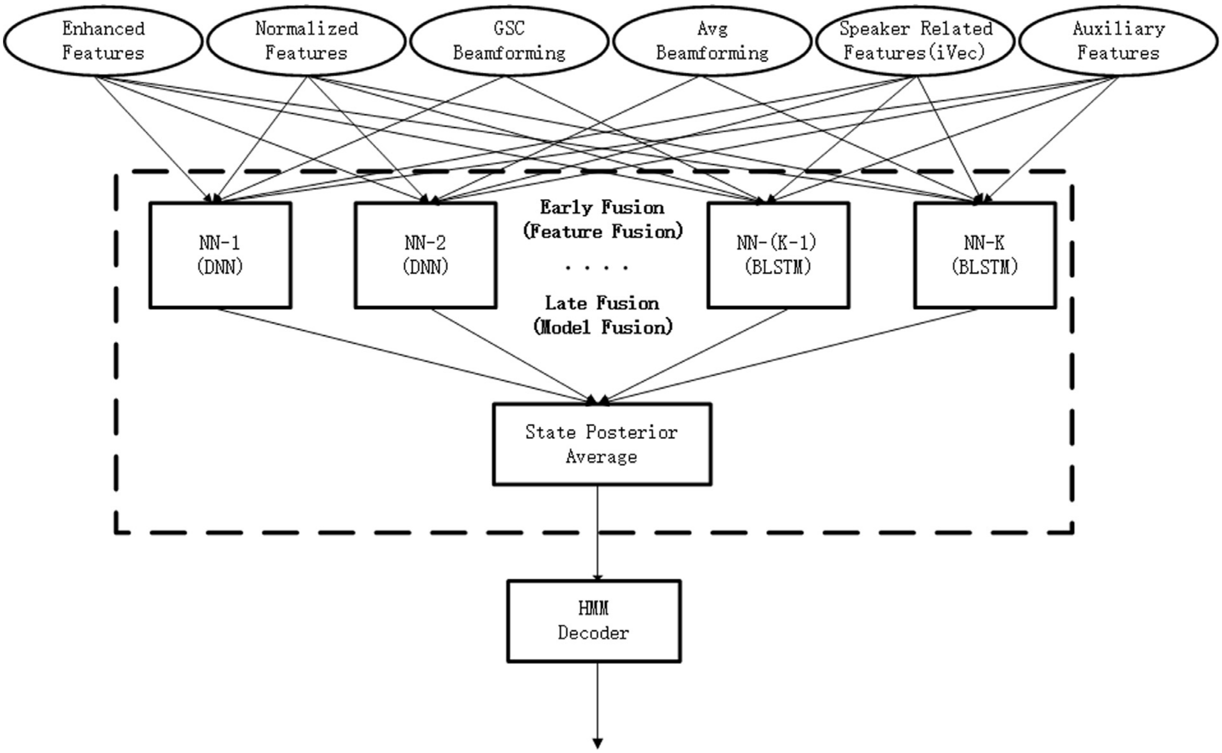


Fig. 2. System overview.

combinations of four speakers each reading 100 utterances in four environments (i.e., 4 × 4 × 100) and 7138 simulated utterances from the WSJ0 training data.

Recordings were made using an array of six Audio-Technica ATR3350 omnidirectional lavalier microphones mounted in holes drilled through a custom-built frame surrounding a Samsung Galaxy tablet computer. The frame is designed to be held in a landscape orientation and has three microphones spaced along both the top and bottom edges, as shown in Fig. 1. All microphones face forward (i.e., towards the speaker holding the tablet), apart from the top-center microphone (Mic 2), which faces backward.

3. System overview

The overall flowchart of our proposed system is illustrated in Fig. 2. The dashed block conceptually denotes a large neural network, consisting of K subnets with different architectures. As for the input, multiple knowledge sources are exploited to generate different feature combinations. Each combination, as an early fusion, includes one type of multi-channel beamforming concatenation, enhanced features, feature normalization, speaker-related features, and auxiliary features, to be elaborated upon in the next section. Each subnet is built independently with different architectures and learning methods. Finally, in recognition, the outputs of the large neural network for each frame are generated by a late fusion of all subnets in the output layer for posterior averaging (Li and Sim, 2013), which are then fed to a decoder with hidden Markov models (HMMs).

4. Multi-channel enhancements

4.1. Delay-and-sum beamforming

Delay-and-sum beamforming is a signal processing technique in which the outputs from an array of microphones are time-delayed so they can be summed. For simplification, we use a simple delay and sum beamforming, namely averaging waveforms, where the delay is set to a constant 0 and the weights to 1/*M*, where *M* is the number of channels.

4.2. MVDR formulation

Given a speech signal $s(t)$ in the target speaker position, the signals received by an array of *M* microphones are time-delayed and amplitude-attenuated versions of $s(t)$ with additional noises and interferences, which can be modeled in the time domain as

$$y_i(t) = g_i s(t - \tau_i) + n_i(t) = x_i(t) + n_i(t) \quad i = 1, 2, ..., M \tag{1}$$

where τ_i is the time of arrival from the speaker location to the *i*th microphone location; g_i is a gain factor to reflect the effects of the propagation energy decay, the amplification gain of the corresponding microphone setting, the directionality of the source and the *i*th microphone; and $x_i(t)$ and $n_i(t)$ are the convolved speech signal and the noise signal received by the *i*th microphone, respectively. In the short-term Fourier transform (STFT) (Mcaulay and Quatieri, 1986) domain, the equation can be expressed as Zhang et al. (2008)

$$\boldsymbol{y}(k,l) = \boldsymbol{g}(k)s(k,l) + \boldsymbol{n}(k,l) = \boldsymbol{x}(k,l) + \boldsymbol{n}(k,l) \tag{2}$$

where *k* is the frequency bin index and *l* is the frame index; $\boldsymbol{x}(k, l)$, $\boldsymbol{s}(k, l)$, and $\boldsymbol{n}(k, l)$ are the complex vectors with *M* dimensions in the STFT domain corresponding to $x_i(t)$, $s(t)$, and $n_i(t)$, respectively; $\boldsymbol{g}(k)$ is the steering vector (Veen and Buckley, 1988). We assume that the analysis window is longer than all of the channel impulse responses and $\boldsymbol{n}(k, l)$ is relatively stationary to be estimated.

The MVDR beamformer applies a set of weights $\boldsymbol{w}(k)$ to the vector $\boldsymbol{y}(k, l)$ such that the variance of the noise component of $\boldsymbol{w}^H(k)\boldsymbol{y}(k, l)$ is minimized, subject to a constraint of unity gain in the target direction,

$$\min_{\boldsymbol{w}} \ \boldsymbol{w}^H(k)\boldsymbol{R_{nn}}(k)\boldsymbol{w}(k), \qquad \text{s.t. } \boldsymbol{w}^H(k)\boldsymbol{g}(k) = 1 \tag{3}$$

where $\boldsymbol{R_{nn}}(k)$ is the spatial correlation matrix of noise and interference defined as the following expectation:

$$\boldsymbol{R_{nn}}(k) = E_l[\boldsymbol{n}(k,l)\boldsymbol{n}^H(k,l)] \tag{4}$$

The closed-form solution of Eq. Eq. (3) is given by Capon (1969):

$$\boldsymbol{w}(k) = \boldsymbol{R_{nn}}^{\frac{-\boldsymbol{1}(k)\boldsymbol{g}(k)}{\boldsymbol{g}^H(k)\boldsymbol{R_{nn}}^{-\boldsymbol{1}(k)\boldsymbol{g}(k)}}} \tag{5}$$

According to Eq. (3), MVDR is a technique that can be used to form an acoustic beam to pick up signals arriving from a direction specified by a steering vector, thereby removing the background noise.

4.3. The proposed MVDR beamforming

The conventional beamformer design obtains the steering vectors by estimating the speaker direction and the microphone array geometry, such as the officially provided beamforming approach (Anguera et al., 2007). In the literature, the robust adaptive beamformers have been extensively studied to deal with direction of arrival (DOA) mismatch, e.g., optimization in the DOA region (Keyi et al., 2005) and the diagonal loading techniques (Zhao et al., 2014). Zhao et al. (2015) designed beamformers robust against microphone gain errors to address the microphone gain estimation problem without any assumptions on the noise field. In Yoshioka et al. (2015), spectral mask-based steering vector estimation without relying on prior information has been introduced. The key of this approach is the unsupervised and accurate estimation of a spectral mask that indicates the presence of speech/nonspeech time–frequency units. In this paper, we directly utilize the first several frames of each test utterance to estimate the noise of the whole test utterance rather than the spectral mask-based approach in Yoshioka et al. (2015).

Supposing that the speech and noise are statistically independent, the spatial correlation matrix of $\boldsymbol{x}(k, l)$, $\boldsymbol{R_{xx}}(k)$ can be estimated as

$$\boldsymbol{R_{xx}}(k) = \boldsymbol{R_{yy}}(k) - \boldsymbol{R_{nn}}(k) \tag{6}$$

where $\boldsymbol{R_{yy}}(k)$ is the spatial correlation matrix of $\boldsymbol{y}(k, l)$. In this study, $\boldsymbol{R_{yy}}(k)$ and $\boldsymbol{R_{nn}}(k)$ are implemented as

$$\begin{aligned}\boldsymbol{R_{yy}}(k) &= \frac{1}{T}\sum_{l=1}^{T}\boldsymbol{y}(k,l)\boldsymbol{y}^H(k,l)\\ \boldsymbol{R_{nn}}(k) &= \frac{1}{T_1}\sum_{l=1}^{T_1}\boldsymbol{n}(k,l)\boldsymbol{n}^H(k,l)\end{aligned} \tag{7}$$

where T is the number of frames of the whole utterance and T_1 (set to 6 in the experiments) is the number of the first several frames of the utterance. Conversely, $\boldsymbol{R_{xx}}(k)$ can be written by definition as

$$\begin{aligned}\boldsymbol{R_{xx}}(k) \quad &= E_l[\boldsymbol{x}(k,l)\boldsymbol{x}^H(k,l)]\\ &= \sigma_S^2(k)\boldsymbol{g}(k)\boldsymbol{g}^H(k)\end{aligned} \tag{8}$$

where $\sigma_s^2(k)$ is the power of the speech signal $s(t)$. Clearly, the positive semi-definite matrix $\boldsymbol{R_{xx}}(k)$ is of rank 1 and the steering vector $\boldsymbol{g}(k)$ can be obtained by computing the principal eigenvector of the estimated $\boldsymbol{R_{xx}}(k)$ (Jones and Ratnam, 2009) from Eq. (6).

4.4. *The generalized sidelobe canceller*

A generalized sidelobe canceller (GSC) (Griffiths and Jim, 1982; Gannot and Cohen, 2004) based on a relative transfer function (Talmon et al., 2009) is adopted in this paper. The GSC as a filter structure can implement a beamformer that minimizes the MVDR objective function, Eq. (3). In this paper, the main difference between the GSC and our proposed MVDR is the estimation of the steering vector. The GSC obtains the steering vector by using DOA estimates, but MVDR obtains the steering vector based on the estimated $\boldsymbol{R_{xx}}(k)$ presented in Section 4.3.

5. Early fusion

5.1. *Beamforming and feature concatenation*

Formulating a strategy to make full use of the multi-channel information of microphone array speech in the neural networks is critical to recognition performance. The existing approaches can be divided into two broad classes: conventional beamforming to generate one single channel output for subsequent processing and channel concatenation. For example, in Liu et al. (2014) and Renals and Swietojanski (2014), the concatenation of the noisy features in each channel of a microphone array outperforms the beamforming approach, especially for moving speech, as it might preserve the signals from all directions. In Li et al. (2014b), the beamformed features concatenated with the noisy features from the main channel of the microphone array yield better recognition performance. In Sainath et al. (2016), the time-domain waveforms are concatenated directly as the input of the CLDNNs, and spatial info can be exploited within the neural net to perform beamforming. However, the microphone array geometry priors are not fully utilized in these approaches.

In our current study, multiple sets of beamforming results are concatenated, as illustrated in Fig. 3. Each beamformed result is generated on a subnet of channels in the microphone array. Multiple beamforming techniques are also adopted for the comparison. One approach is the waveform averaging of the specified channels, denoted as *Avg* in Table 1. Another approach is a generalized sidelobe canceller (GSC). The last approach is a simplified version of the MVDR beamformer from Yoshioka et al. (2015), which achieves the best recognition performance in comparison to the other two. A detailed formulation and description of the MVDR approach will be given in the next subsection. After beamforming, multiple sets of features, such as log Mel-filterbank (LMFB) features, speaker-related features, and other auxiliary features extracted from the enhanced signals, will be elaborated upon in the following subsections. These features are exploited to generate different combinations for the input to each subnet, but the acoustic context (the number of neighboring frames) of each concatenated feature set is different according to its believed importance to acoustic modeling.

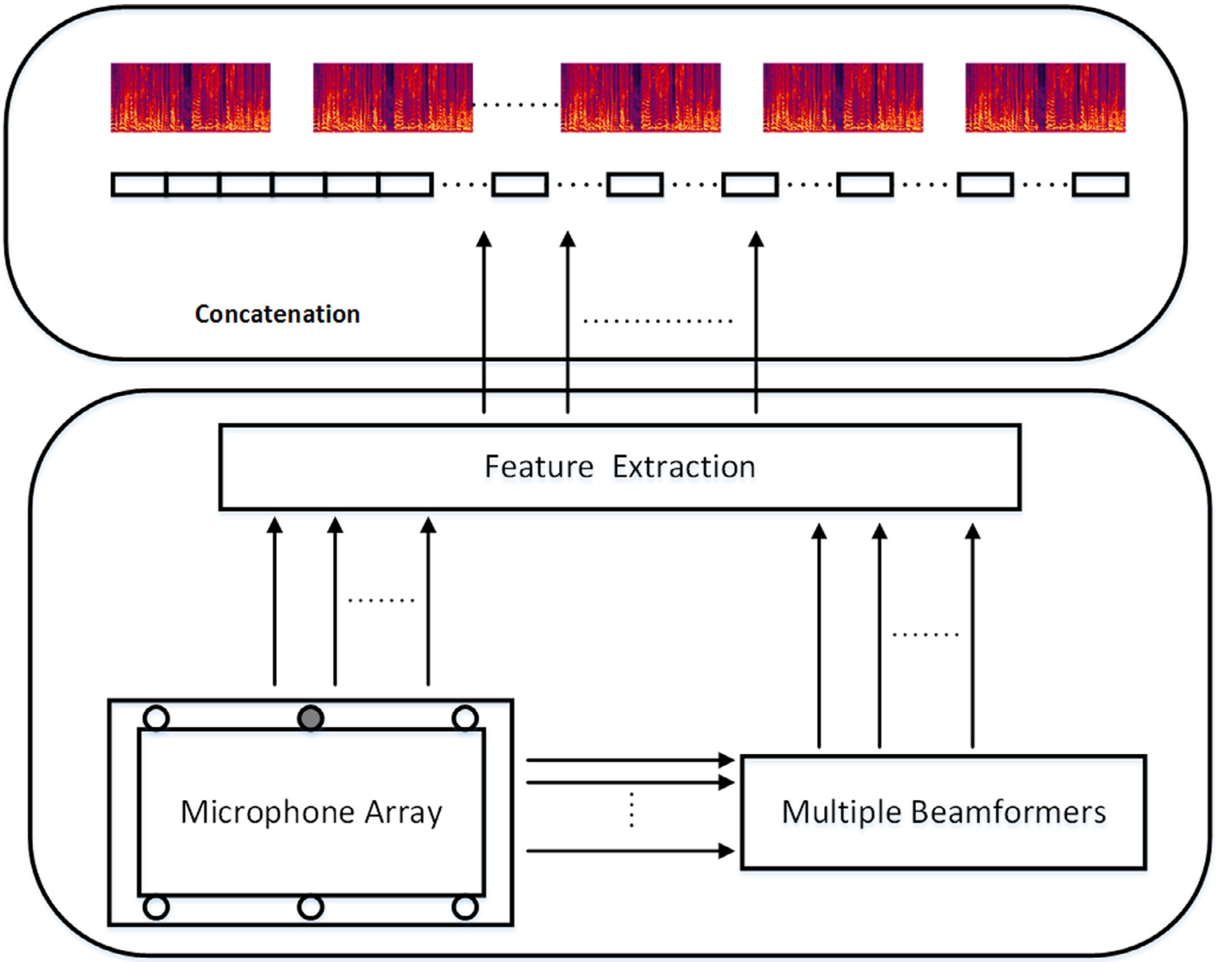


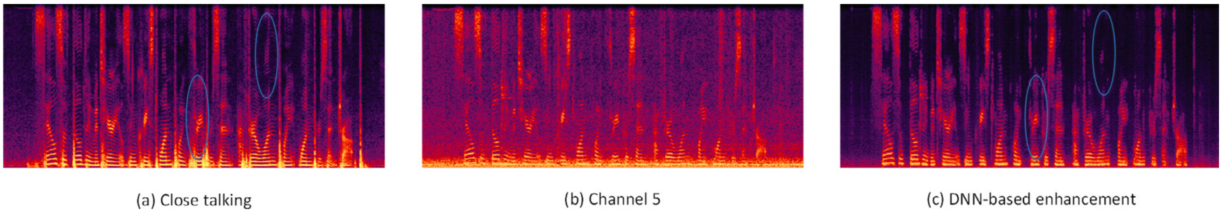
Fig. 3. Feature concatenation with multiple beamformers.

5.2. Enhanced features

To demonstrate the effectiveness of the enhanced features (denoted as **Enh**) combined with the beamforming concatenation, we used the officially provided beamforming approaches (Barker et al., 2015). The source localization technique in Loesch and Yang (2010) was used to track the target speaker, and the speech signal was estimated by time-varying MVDR beamforming with a diagonal loading (Mestre and Lagunas, 2003) architecture so we can

Table 1
Description of the different concatenated features.

Feature	Frames	Channels	Beamformers	Type	Dimensions
Avg1	11	4,5,6	Averaging	fbank+pitch	1386
Avg2	1	1,3	Averaging	fbank+pitch	126
CH2	1	4,5,6	No	fbank+pitch	126
Enh	1	1,3,4,5,6	Official	fbank+pitch	126
GSC1	11	4,5	GSC	fbank+pitch	1386
GSC2	1	5,6	GSC	fbank+pitch	126
GSC3	1	1,3	GSC	fbank+pitch	126
iVec1	1	4,5,6	No	ivector	20
iVec2	1	1,3	No	ivector	20
CG1	1	4,5,6	No	cochleagram	30
CG2	1	1,3	No	cochleagram	30


(a) Close talking

(b) Channel 5

(c) DNN-based enhancement

Fig. 4. Illustration of spectrograms for (a) an utterance recorded by a close-talking microphone (target speech) (b) the utterance from channel 5 of the real data. (c) DNN-based enhanced speech.

directly concatenate different types of features as our input. In the following, we introduce some of the robust features used in our work.

In our recent work (Xu et al., 2015; Tu et al., 2015; Du et al., 2014), the DNN-based single-channel enhancement approach has been proven to be effective in single-channel speech recognition for simulation data. We now investigate its effectiveness on the real data of the CHiME-3 task. In this study, the DNN as a regression model is used to predict the log-power spectral (LPS) features of clean speech given the input LPS features of noisy speech with an acoustic context, which can be regarded as a pre-processing technique. The noisy speech is selected from channel 5 of the microphone array. Overall, 7138 utterance pairs of simulated data and 1600 utterance pairs of real data are used for training. However, the DNN pre-processing system could not yield a performance gain over the unprocessed system, as confirmed in Hori et al. (2015).

To explain this, we show the spectrograms of an utterance example illustrated in Fig. 4. Fig. 4(a) is a spectrogram of an utterance recorded by a close-talking microphone, which can be considered as the target clean speech in the real data. Fig. 4(b) is a spectrogram of the utterance from channel 5 of the real data. By the comparison between Fig. 4(a) and (c), severe speech distortions can be observed in the marked elliptical areas, which lead to incorrect recognition results. This problem is possibly due to (i) the limited real training data; (ii) the target speech of the real data still being noisy, as shown in Fig. 4(a); and (iii) that only four speakers were included in the training set. Based on this analysis, the DNN pre-processing technique is not further investigated in this study.

5.3. Normalized features

Utterance-based feature normalization is widely used in ASR systems to reduce the effect of possible irrelevant variabilities due to speaker, background noises and channel distortions. Two normalization approaches, namely, mean normalization (denoted as *MN* in Table 2) and mean variance normalization (denoted as *MVN* in Table 2), are applied to the acoustic features. MVN is more effective for additive noises, especially with low SNRs, while MN is more stable for the high-SNR cases.

Table 2
Description of 12 subsystems (DAM denotes (D)NN using (A)verage beamforming and (M)ean normalization, and BGV denotes (B)LSTM using (G)SC beamforming and mean (V)ariance normalization).

	Feature fusion	NN type	Parameters (*M*)	One iteration (*h*)
DAM	MN(Avg1+Avg2+CH2+Enh)+iVec1+iVec2	DNN	112	0.6
DGM	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	DNN	112	0.6
DAV	MVN(Avg1+Avg2+CH2+Enh+CG1+CG2)+iVec1+iVec2	DNN	112	0.6
DGV	MVN(GSC1+GSC2+GSC3+Enh+CG1+CG2)+iVec1+iVec2	DNN	112	0.6
LAM	MN(Avg1+Avg2+CH2+Enh)+iVec1+iVec2	LSTM-RNN	116	1.5
LGM	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	LSTM-RNN	116	1.5
LAV	MVN(Avg1+Avg2+CH2+Enh)+iVec1+iVec2	LSTM-RNN	116	1.5
LGV	MVN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	LSTM-RNN	116	1.5
BAM	MN(Avg1+Avg2+CH2+Enh)+iVec1+iVec2	BLSTM-RNN	222	2.3
BGM	MN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	BLSTM-RNN	222	2.3
BAV	MVN(Avg1+Avg2+CH2+Enh)+iVec1+iVec2	BLSTM-RNN	222	2.3
BGV	MVN(GSC1+GSC2+GSC3+Enh)+iVec1+iVec2	BLSTM-RNN	222	2.3

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004

5.4. Speaker-related Features

Similar to Saon et al. (2013), the i-vectors (denoted as *iVec* in Table 1) that represent the speaker information are extracted via the standard procedure (Dehak et al., 2011; Glembek et al., 2011) as parallel features fed to the input layer of neural nets. The main idea is that the speaker- and channel-dependent Gaussian mixture model (GMM) supervector $\boldsymbol{s}$ can be formulated as

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{9}$$

where $\boldsymbol{m}$ is the mean supervector of the universal background model (UBM), $\boldsymbol{T}$ is a low-rank matrix representing M bases spanning the subspace with important variabilities in the mean supervector space, and $\boldsymbol{w}$ is a standard normal distributed vector of size M. The i-vector is the maximum a posteriori (MAP) point estimation of $\boldsymbol{w}$ given the speech segments. The main advantage of the i-vector based speaker adaptation approach is that the architecture of the neural net remains unchanged, so it is unnecessary to perform the first-pass decoding. Inspired by the beamforming concatenation, the multi-channel i-vectors are extracted corresponding to each beamforming result, and they are verified more effectively than the single-channel i-vector. Note that for both training and testing, the i-vector is estimated based on the utterances of a single speaker and only changed across different speakers.

5.5. Auxiliary features

Besides the commonly used LMFB features, other auxiliary features are also adopted. One feature set is the pitch and probability-of-voicing features proposed in Ghahremani et al. (2014) and Metze et al. (2013), which are tuned for the ASR systems. It is believed that those features not only give large improvements for tonal language recognition but also yield remarkable gains for non-tonal languages, which is confirmed in our task. The other set is the cochleagram (CG) features that are well verified for ASR (Chen et al., 2014). In our experiments, the pitch-related features (Ghahremani et al., 2014) are always concatenated with the LMFB features in each system of Table 1, whereas the CG features are optionally used.

As mentioned above, in early fusion, diverse features are concatenated together. One issue is to control the input feature dimension to avoid possible performance degradation. Suppose that the dimension of the basic acoustic features is D_1 and the size of the acoustic context is τ frames. The number of channels after beamforming is N. The dimensions of the i-vector and auxiliary features are D_2 and D_3, respectively. Then, the final dimension for the input feature vector is $D_1 * \tau + N * D_1 + N * D_2 + N * D_3$, which means that the acoustic context expansion is only applied to the main channel of the basic acoustic features.

6. Late fusion

6.1. Acoustic modeling

Three types of neural nets are adopted as subnets: DNN, long short-term memory (LSTM)-based RNN (Sak et al., 2014a), and bi-directional LSTM (BLSTM)-based RNN (Graves et al., 2013). Before the neural network training, state labels should be generated using forced alignment via a state-of-the-art system with GMM-HMMs (Tachioka et al., 2013). The only difference is the use of the multi-channel concatenation of acoustic features after the waveform average beamforming. This set of state labels is used for the training of all subnets. For the DNN training, the Kaldi recipe for the CHiME-2 challenge (Weng et al., 2014) is adopted with the standard procedure, namely, pretraining using restricted Boltzmann machines plus cross-entropy (CE) training. The DNN can be refined by re-alignment (ReFA) and sequence discriminative training using the state-level minimum Bayes risk (sMBR) criterion (Vesel et al., 2013). For training the LSTM-RNN and BLSTM-RNN, the CE (Sak et al., 2014a) and sMBR criterion (Sak et al., 2014b) are adopted with the truncated backpropagation through time (BPTT) learning algorithm to update the model parameters. The neural network types, parameters and computational speed are shown in Table 2. The first letter of each subsystem abbreviation represents the neural network type (D, L, and B denote (D)NN, (L)STM, and (B)LSTM, respectively), the second letter represents the beamforming type (A and G denote (A)veraging beamforming and (G)SC beamforming, respectively), and the last letter represents the feature normalization (M and V denote (M)ean normalization and mean (V)ariance normalization, respectively).

Table 3
WERs (in %) of GMM-based systems trained with different channels on the development set of real data.

System	%WER for real dev.				
	BUS	CAF	PED	STR	AVG.
CH1	25.25	27.17	17.33	22.96	23.18
CH2	64.02	56.17	51.03	70.87	60.52
CH3	27.35	27.82	16.99	22.78	23.74
CH4	26.64	19.32	16.98	22.46	21.35
CH5	25.93	17.46	13.13	17.89	**18.6**
CH6	24.07	19.81	14.88	18.63	19.35

6.2. *Language modeling*

In this paper, in addition to the originally provided 3-gram language model (LM) as the baseline, a RNN LM (Mikolov et al., 2010) and a 5-gram LM with modified Kneser−Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1996) are generated using the WSJ0 text corpus. The RNN LM is more effective and is composed of a neural network including a hidden layer with re-entrant connections to itself with a one-word delay. The activations of the hidden units play the role of memory, keeping a history from the beginning of the speech. Accordingly, the RNN LM can robustly estimate word probability distributions by representing the histories smoothed in the continuous space and taking long-distance inter-word dependencies into account. Mikolov et al. reported that the RNN LM yielded a large improvement in recognition accuracy when combined with a standard *n*-gram model (Mikolov et al., 2010). In the decoding phase, word lattices are first generated using the baseline LM, namely, the standard 5k WSJ 3-gram with entropy pruning. Then, *N*-best lists are generated from the lattices using the 5-gram LM. Finally, the *N*-best lists are re-ranked using a linear combination of the 5-gram and RNN LMs. The best-ranked hypothesis is selected as the recognition result of each single system.

6.3. *System combination*

For all *K* subnets, the outputs share the same tied state set from the HMM topology of the GMM-HMM or DNN-HMM system. Thus, late fusion can be implemented by a simple strategy of state posterior averaging in the output layers (Li and Sim, 2013). This approach has been verified to be more effective than lattice fusion or ROVER (Fiscus, 1997) in our experiment, which is reasonable as fusion at the frame level (state level) is done at a higher resolution than that at the text level and is not affected by the language model. Back to Fig. 2, if we treat early fusion and late fusion as internal operations of the large neural net in the dashed box, then the input might be a high-dimensional vector with diverse features from multiple knowledge sources, while the output is still the normal-state posterior representation.

7. Experiments and results

We evaluate the ASR performance of the proposed concatenation methods on the multi-condition training schemes for the CHiME-3 task and compare them to the baseline ASR systems provided by the challenge sponsors. Both the GMM and DNN acoustic models were used in the ASR systems.

7.1. *Empirical experiments on beamforming and concatenation*

Due to space limitations in Du et al. (2015), the feature concatenation configuration was directly given there without explanation. Here, we detail our proposed approach in multi-channel scenarios. The following experiments show the importance of a reasonable concatenation and a full usage of data from all channels.

First, Table 3 compares models trained on 6-channel data sets in terms of the WER on the development set of real data. We adopted a GMM-HMM system using a 91-dimensional feature vector, consisting of 13-dimensional

Table 4
WERs (in %) of GMM-based systems trained with different beamformings on the development set of real data.

System	%WER for real dev.				
	BUS	CAF	PED	STR	AVG.
CH5	25.93	17.46	13.13	17.89	18.60
Enh	21.63	17.67	16.45	18.20	18.49
AvgAll	28.50	25.47	16.52	21.90	23.1
AvgNoch2	24.30	18.13	13.97	20.23	19.16
Avg1	24.41	16.52	12.29	16.13	**17.34**
Avg2	22.70	22.89	14.29	18.82	19.68

Mel-frequency cepstral coefficients (MFCCs) with a 7-frame context expansion. The result confirmed that the quality of channel 5 was the best among the 6 channels, and it would play a key role in beamforming and feature concatenation in the subsequent experiments.

Second, Table 4 gives a WER comparison with Avg beamforming of different channels on the development sets of real data using the above GMM-HMM acoustic model. For all beamforming, Channel 5 was used due to its reliable speech quality. “CH5” and “Enh” denote the acoustic model trained on the speech data of channel 5 and the enhanced data by the official tools using 6-channel data, respectively. For the waveform averaging beamforming, “AvgAll” , “AvgNoch2”, “Avg1’, and “Avg2” denote the averages of channels (1,2,3,4,5,6), (1,3,4,5,6), (4,5,6), and (1,3), respectively. When training and testing data were both processed by the provided officially enhanced data, the WER of 18.49% for the real data was slightly decreased compared to that of the single channel baseline (CH5) of 18.60%. This result shows a weakness of the conventional multi-channel enhancement for ASR. One reason that the “AvgAll” system performed the worst at 23.1% may be that channel 2 contains a lot of noise. By excluding channel 2, we found that the Avg1 system at 17.34% outperformed the baseline system by a large margin, which indicates that the performance of waveform averaging beamforming depends upon the speech quality of each channel, and averaging beamforming is more robust to speaker motion than the official beamformer.

The above experiments were conducted using the GMM-HMM model for a quick verification of our point. Next, the DNN-HMM model was adopted. Table 5 shows the performance of our proposed simple concatenation system on the development sets of real data. In the DNN-based baseline system, 11 frames of 40-dimension LMFB features with their first-order and second-order derivatives were used, resulting in 1320-dimensional features. The DNN has 7 hidden layers with 2048 hidden units at each layer and the 1965-dimensional softmax output layer, corresponding to the senones of the GMM-HMM system. Compared to the Avg1 results in Table 4, we found that the accuracy of DNN was better than that of GMM for the Avg1 system. In the experiments to follow here, the DNN system setting was adopted. As for the simple concatenation system, the frames of each concatenated feature were the same, so the input dimensions were an integral multiple of 1320, namely, 1320 × N, where N is the number of concatenated features. The first two rows of Table 5 show that our proposed beamforming and simple concatenation “Avg1+Avg2” system consistently outperformed the “Avg1” system for all testing cases, e.g., a relative WER reduction of 6.93% was achieved for the development sets on average. The result indicates that beamforming on different channels might be strongly complementary to make full use of multi-channel information. Finally, channel 2 containing much

Table 5
WER (in %) comparison at different simple concatenations of the Avg beamforming of different channels on the development sets of real data using the DNN acoustic model.

System	Feature dimension	%WER for real dev.				
		BUS	CAF	PED	STR	AVG.
Avg1	1320	22.48	14.65	12.13	16.50	16.44
Avg1+Avg2	2640	20.52	14.50	11.12	15.06	**15.3**
Avg1+Avg2+CH2	3960	24.78	18.08	12.86	16.56	18.07

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004

Table 6
WERs (in %) for feature concatenations with different frame settings on the development sets of real data using the DNN acoustic model.

System	Feature dimension	%WER for real dev.				
		BUS	CAF	PED	STR	AVG.
Avg1+Avg2(11+11)	2772	15.93	12.30	7.98	11.47	11.92
Avg1+Avg2(11+5)	2016	16.15	11.67	7.86	11.74	11.86
Avg1+Avg2(11+3)	1764	16.60	11.28	7.76	11.49	11.78
Avg1+Avg2(11+1)	1512	16.74	11.83	8.79	11.58	11.98

noise was used as the noise source to improve the noise robustness of ASR, for example, noise-aware training. However, no gains were observed by using the data from channel 2. This result may be due to size of the input dimension, which was 3960 and was much larger than the number of nodes of the hidden layer, resulting in an information loss when the input features were transferred into the first hidden layer.

Then, we concatenated the 2-dimensional pitch features (pitch and probability-of-voicing) mentioned in Section 5.5 into the 40-dimensional LMFB features and applied utterance-based MN plus their first-order and second-order derivatives, resulting in a 126-dimensional feature vector at each frame. It was clearly seen that MN consistently reduced the WER across all test cases when comparing the second row result of 15.3% via 2640-dimensional features in Table 5 with the first row result of 11.92% via 2772-dimensional features in Table 6. MN was able to reduce the channel mismatches and noise effects on the features.

Finally, Table 6 shows that the WER is basically invariable with the changing frames of Avg2, which means that more complementary features can be concatenated as the input dimension is reduced. The “Avg1+Avg2(11+11)” system denotes that the frames of the concatenated Avg1 and Avg2 features were both 11. Although the “Avg1+Avg2(11+3)” system slightly outperformed the “Avg1+Avg2(11+1)” system, the dimension of the latter was smaller. Thus, the settings of the system “Avg1+Avg2(11+1)” were adopted and denoted as “Avg1+Avg2” for convenience in the following experiments, and the number of concatenated features is one, except for Avg1, which has 11 frames.

7.2. *Feature concatenation: early fusion*

In this section, experiments on early fusion were reported. Table 7 gives a WER comparison at different stages of early fusion for the DAM system on the development and test sets of real data. Our proposed beamforming and concatenation system “Avg1+Avg2” consistently outperformed the baseline system for all testing cases, e.g., relative

Table 7
WER (in %) comparison of different early fusions for the DAM system on the development and test sets of real data.

System	Feature dimension	BUS	CAF	PED	STR	AVG.
	Development set of real data					
CH5	1386	20.93	12.89	9.18	13.58	14.15
Avg1+Avg2	1512	16.74	11.83	7.79	11.58	11.98
+CH2	1638	15.74	11.52	7.92	11.47	11.66
+Enh	1764	14.35	10.06	7.76	10.50	10.67
+iVec1+iVec2	1804	12.33	9.45	6.83	10.37	9.75
+ReFA	1804	11.70	9.00	6.86	9.76	9.33
+sMBR	1804	10.87	7.92	6.14	8.88	8.45
	Test set of real data					
CH5	1386	34.77	26.24	20.76	16.23	24.50
Avg1+Avg2	1512	28.04	22.23	17.30	13.54	20.28
+CH2	1638	27.28	21.22	17.28	12.87	19.66
+Enh	1764	25.31	21.26	15.88	12.27	18.68
+iVec1+iVec2	1804	22.79	20.02	15.13	11.60	17.39
+ReFA	1804	21.33	18.88	14.78	11.32	16.58
+sMBR	1804	19.09	16.74	13.19	10.53	14.89

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004

Table 8
WER (%) comparison of different early fusions for the DGM system on the development and test sets of real data.

System	Feature dimension	BUS	CAF	PED	STR	AVG.
		Development set of real data				
GSC1	1386	16.92	10.44	7.57	11.28	11.55
+GSC2	1512	16.37	9.73	7.23	10.57	10.98
+GSC3	1638	15.96	9.97	7.26	10.60	10.95
+Enh	1764	14.62	9.29	7.23	9.97	10.28
+iVec1+iVec2	1804	13.57	8.83	6.90	10.09	9.85
+ReFA	1804	13.39	8.45	6.90	9.95	9.68
+sMBR	1804	12.16	8.14	6.12	8.64	8.77
		Test set of real data				
GSC1	1386	27.16	21.78	17.12	13.34	19.85
+GSC2	1512	25.82	20.94	15.98	12.83	18.89
+GSC3	1638	25.69	20.86	15.21	12.63	18.59
+Enh	1764	24.27	21.22	15.56	12.29	18.33
+iVec1+iVec2	1804	22.36	20.30	14.52	11.58	17.19
+ReFA	1804	22.56	19.91	14.87	11.60	17.24
+sMBR	1804	20.02	17.26	12.91	10.40	15.15

WER reductions of 15.3% and 17.2% were achieved for the development and test sets on average. Then, by appending the channel 2 features, the recognition performance was slightly improved, in contrast to the point that the high dimension of input concatenated features may degrade the performance in Section 7.1. More interestingly, the concatenation with the enhanced features brought about an absolute 1% WER reduction for both the development and test sets. However, according to Table 4, no obvious improvements were observed by the use of the enhanced features only. This indicates the necessity of the parallel beamformed and enhanced features, which might be strongly complementary. Furthermore, the additional i-vector features gave remarkable gains, demonstrating the effectiveness of these speaker-adapted features. As for DNN training, ReFA and sMBR could consistently reduce the WER. Overall, relative WER reductions of 40.3% and 39.2% were yielded from the baseline system for the development and test sets, respectively. By considering that the test set was more difficult than the development set, these similarly relative improvements show a generalization ability of our proposed early fusion strategy. The above results showed that the more complementary features that could be concatenated, the better the performance improvements are.

Table 8 shows a WER comparison at different stages of early fusion for the DGM system on the development and test sets of real data. Similar observations to those for DAM as in Table 7 could be made. The main difference between DAM and DGM was the use of GSC-based beamforming. It is interesting that in the stage of pure beamforming concatenation, the GSC-based approach outperformed the waveform average approach, e.g., with the average WER decreasing from 19.66% to 18.59% on the test set. However, by a performance comparison of the final systems, DAM and DGM, we could make the opposite observation, that waveform averaging was slightly better than that of GSC, which implies that the simple average operation in the time domain was a robust beamforming approach. Finally, both DAM and DGM gave significant gains over the baseline system, and each feature set in the early fusion stage made a contribution to reducing the WERs.

7.3. System combination: late fusion

Before presenting the late fusion results, the recognition performances of the 12 subsystems to be combined are shown in Table 9. Clearly, no single subsystem could achieve the best performance in all environments. For the test set with real data, DAM achieves the best performance on average, but not the best for the PED and STR environments. There were 7 subsystems with at least one best performance case. Those observations deliver important messages. On the one hand, the noise statistics should be quite different in the four test environments, so that each subsystem with one feature combination could not optimally handle all of the noise conditions. On the other hand, all of the subsystems might be complementary, which was one key motivation for our proposed late fusion strategy.

Table 9
WER (in %) comparison of 12 subsystems on the development and test sets of real and simulated data.

			DAM	DGM	DAV	DGV	LAM	LGM	LAV	LGV	BAM	BGM	BAV	BGV
Dev	Simu	BUS	6.64	5.91	7.24	6.52	7.33	6.80	7.27	6.50	6.83	**5.80**	6.65	6.27
		CAF	**9.09**	9.23	10.27	10.04	10.18	10.41	9.41	10.03	9.73	9.20	9.60	9.31
		PED	6.17	**5.72**	6.90	6.39	6.78	6.42	7.02	6.65	6.45	6.09	6.70	6.25
		STR	8.33	6.96	9.01	7.29	8.76	8.16	8.91	7.57	7.86	**6.90**	8.24	7.52
		Avg.	7.56	**6.96**	8.36	7.56	8.26	7.95	8.15	7.69	7.72	7.00	7.80	7.34
	Real	BUS	10.87	12.16	12.16	**7.08**	12.94	14.43	13.26	13.69	11.11	12.45	12.82	12.17
		CAF	**7.92**	8.14	8.83	10.12	10.29	9.75	9.62	9.73	8.86	8.63	8.66	8.81
		PED	6.14	6.12	**5.94**	8.48	7.48	7.80	7.77	8.16	7.05	7.17	7.02	6.80
		STR	8.88	8.64	**7.47**	10.31	10.22	10.50	9.84	10.29	9.44	9.05	8.95	8.98
		Avg.	**8.45**	8.77	9.10	9.00	10.23	10.62	10.12	10.47	9.12	9.33	9.36	9.19
Test	Simu	BUS	7.68	6.74	7.45	13.32	8.03	7.71	6.82	6.71	7.49	7.27	6.93	**6.37**
		CAF	11.60	9.51	11.00	9.17	12.07	12.03	**9.08**	9.88	10.66	10.96	9.69	9.56
		PED	11.58	**8.37**	10.96	6.37	11.45	10.70	9.56	8.87	9.77	9.77	9.23	8.74
		STR	11.52	**9.21**	11.94	8.98	12.85	10.89	11.21	10.09	11.62	9.97	11.04	9.60
		Avg.	10.59	**8.46**	10.34	9.46	11.10	10.33	9.17	8.89	9.89	9.49	9.22	8.57
	Real	BUS	**19.09**	20.02	23.35	25.24	22.08	23.72	22.21	22.43	20.38	22.99	21.24	21.07
		CAF	**16.74**	17.26	19.78	20.58	19.01	19.74	18.85	19.44	17.02	17.59	17.52	17.69
		PED	13.19	12.91	14.69	14.07	15.19	16.42	14.28	15.53	13.85	14.41	13.83	**12.65**
		STR	10.53	10.40	11.90	12.57	11.34	12.18	11.49	11.09	10.65	10.96	11.28	**10.16**
		Avg.	**14.89**	15.15	17.43	18.11	16.90	18.02	16.70	17.12	15.47	16.49	15.97	15.39

Table 9 illustrates a WER comparison of different combinations in late fusion on the development and test sets of the real and simulated data. We designed the fusion experiments from two aspects, namely, the fusion of different neural networks with a fixed input feature combination and the fusion of different inputs with a fixed type of neural network. For descriptive simplification, we rename the 12 systems (DAM, DGM, DAV, DGV, LAM, LGM, LAV, LGV, BAM, BGM, BAV, BGV) as S1–S12. From the results of F(1,5,9), F(2,6,10), F(3,7,11), F(4,8,12), significant improvements were achieved by fusing different architectures (DNN, LSTM-RNN, BLSTM-RNN), e.g., WER on the real test data was reduced from 14.89% in the best single subsystem to 12.1% in F(3,7,11) on average, indicating that different architectures could help each other in predicting the state posteriors at the output layer (Table 10). With the fixed neural network type, the improvements by fusing different feature inputs were also significant on the real data. The small dynamic range of the WERs for the first 7 fusion systems is interesting. By fusing all 12 subsystems

Table 10
WERs (in %) of different system combinations on the development and test sets of real and simulated data (F(1,5,9) means fusion of subsystems 1, 5, and 9).

			F(1,5,9)	F(2,6,10)	F(3,7,11)	F(4,8,12)	F(1–4)	F(5–8)	F(9–12)	F(1–12)
Dev	Simu	BUS	5.53	4.97	5.38	**4.82**	5.50	5.16	5.18	5.07
		CAF	7.98	7.96	7.61	8.05	8.10	8.17	8.01	**6.95**
		PED	5.41	5.22	5.44	5.18	4.99	5.49	5.49	**4.73**
		STR	6.61	5.80	6.36	5.97	6.11	5.87	6.05	**5.62**
		Avg.	6.38	5.99	6.2	6.01	6.17	6.17	6.18	**5.59**
	Real	BUS	9.20	10.33	9.81	10.86	10.05	10.31	10.08	**8.76**
		CAF	7.26	7.27	7.02	7.17	7.39	7.37	7.01	**6.37**
		PED	5.66	6.02	5.68	5.99	5.40	5.74	6.00	**5.03**
		STR	7.46	8.23	7.27	7.92	7.71	7.26	7.34	**6.44**
		Avg.	7.4	7.96	7.44	7.89	7.64	7.67	7.61	**6.65**
Test	Simu	BUS	5.88	5.83	5.49	5.70	6.33	6.11	5.79	**5.30**
		CAF	9.25	9.11	7.86	8.03	8.52	8.91	8.46	**7.71**
		PED	8.69	8.07	7.38	6.99	8.11	7.94	7.60	**6.82**
		STR	8.91	**7.77**	9.02	8.03	9.00	8.31	8.74	7.96
		Avg.	8.19	7.70	7.44	7.20	7.99	7.82	7.65	**6.95**
	Real	BUS	15.87	17.74	16.04	17.46	16.79	17.76	16.30	**13.78**
		CAF	13.00	13.52	13.35	14.21	14.34	14.51	13.99	**11.36**
		PED	11.53	11.53	10.69	10.63	10.71	12.31	11.15	**9.30**
		STR	8.50	9.10	8.33	9.02	9.17	9.34	9.02	**7.77**
		Avg.	12.22	12.97	12.10	12.83	12.75	13.48	12.62	**10.55**

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004

Table 11
WER (in %) comparison between different beamformers on the test sets of real data using the CH5 and retrained acoustic models.

Training data	Test data	%WER for real test				
		BUS	CAF	PED	STR	AVG.
CH5	CH5	34.77	26.24	20.76	16.23	24.50
CH5	Avg1	37.03	24.04	17.86	14.98	23.48
CH5	GSC1	36.19	22.38	18.25	14.44	22.81
CH5	MVDR	21.87	17.05	14.18	12.63	**16.43**
CH5+MVDR	MVDR	18.93	16.02	13.59	12.23	**15.19**

in F(1−12), a relative WER reduction of 29.1% (from 14.89% to 10.55%) was obtained from the best single subsystem on the real test data. The F(1−12) system consistently achieves the best results for both the development and test sets of real data, and this observation could be extended to most of the best systems with the simulation data.

7.4. Train-test beamforming mismatch

Next, we further enhanced our proposed fusion strategies by improved beamforming on the test sets of the real data using the CH5 acoustic model. We find that our simplified MVDR beamformer, which used the data from all 6 channels, could achieve a relative WER reduction of 32.94% (from 24.50% in the third row to 16.43% in the bottom row) without any acoustic model retraining in Table 11. The WER on the real test data is reduced from 16.43% in the model that was trained on CH5 data to 15.19% in the model that was retrained using CH5+enhanced data. Hence, the retraining can bring about a great improvement.

Then, we directly used the proposed MVDR beamformer to replace the other beamformed features in the test stage, and the acoustic models of the systems presented in Table 2 remained unchanged. Table 12 shows a WER comparison with the replacements of the different beamformed features for systems DAM, DAV, DGM, and DGV on the test set of real data. For systems DAM and DAV, the features of Avg1 and Avg2 were replaced with the MVDR and Avg1, respectively, and we denoted the two systems as improved DAM and DAV, respectively. Compared to the DAM, with a WER result of 14.89% in the bottom row of Table 9, the concatenation of the MVDR beamformed features remarkably reduced the WER to 11.68% for improved DAM, representing a relative WER reduction of 21.56% on average, although there existed some mismatching between the concatenated features of the training and test stages. The result shows that our training data can be enlarged with different concatenated features in the limitation of the training data. The last two rows of Table 12 also showed similar WER reductions in the GSC systems.

7.5. Language model rescoring

Finally, a large-scale language model (Hori et al., 2015) was adopted to further improve the ASR performance. First, we used the WSJ0 text corpus to train a 5-gram LM and a RNN LM, which was a class-based LM with 200 word classes and 500 hidden units. Then, the 5-gram and RNN LM probabilities were linearly combined, and the best weights of the combination were chosen depending on the development set. The performance of the re-scoring

Table 12
WER (in %) comparison of the four improved DNN systems on the test set of real data using our proposed MVDR beamforming in the test stage.

Improved system	Test data	%WER for real test				
		BUS	CAF	PED	STR	AVG.
DAM	MVDR + Avg2+	12.9	13.00	11.08	9.75	11.68
DAV	CH2+Enhan+iVec	13.87	13.80	10.72	10.63	12.26
DGM	MVDR+GSC1+	15.63	15.41	12.93	11.52	13.87
DGV	CH2+Enhan+iVec	16.02	15.13	11.70	10.83	13.42

Table 13
WER (in %) comparison of the 5-gram and RNN LMs for the improved DAM system on the test set of real data.

LM	%WER for real test				
	BUS	CAF	PED	STR	AVG.
3-gram	12.9	13.00	11.08	9.75	11.68
5-gram	10.67	11.23	10.11	8.72	10.18
+ RNN LM	9.53	9.69	8.84	8.26	9.08

Table 14
WER (in %) comparison of 3-gram, 5-gram and RNN LMs for fusing the four improved systems (DAM, DAV, DGM, and DGV) on the test set of real data.

LM	%WER for real test				
	BUS	CAF	PED	STR	AVG.
3-gram	12.45	12.25	9.70	9.21	10.90
5-gram	10.58	10.48	8.43	8.44	9.48
+ RNN LM	9.18	9.28	7.51	7.60	8.39

of the hypothesized word lattice using a RNN LM for the improved DAM system mentioned in the above section is shown in Table 13. We can observe that both the 5-gram and RNN LMs further improve the performance over the 3-gram model for the improved system DAM in Table 12. The improved DAM system achieved a WER of 9.08%, which is the best DNN system with one-pass decoding among all of the systems submitted to CHiME-3. The fusion results of the four improved systems (DAM, DAV, DGM, DGV) shown in Table 12 were also provided in Table 14, and the best WER of 8.39% shown in the bottom row is much better than those of the Top-2 (a WER of 9.10%) (Hori et al., 2015) and Top-3 (a WER of 10.55%) (Du et al., 2015) systems submitted to CHiME-3.

8. Conclusion and future work

In this paper, we propose to integrate multiple knowledge sources denoted by multiple feature sets into deep neural networks with different architectures. The proposed early fusion is adopted for local feature concatenation to deal with incomplete features caused by imperfect beamforming, while the proposed late fusion acts as a model average of complementary systems. Since improved beamforming (the simplified version of the MVDR beamformer in Yoshioka et al., 2015), enhanced features (as demonstrated in pitch features Ghahremani et al., 2014, speaker-adapted features (Saon et al., 2013) and normalized features) and powerful language models (as demonstrated in RNN LM Mikolov et al., 2010) are also available, we incorporate them for improved early fusion and late fusion. By replacing the weak beamformers, such as Avg and GSC, without changing the trained system, a huge improvement can also be obtained. This can also save much retraining time when a better beamformer is provided that can be directly used. In future work, we will expand our structure to any microphone array and design a structure to automatically achieve the optimal concatenation.

References

Anguera, X., Wooters, C., Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. 15 (7), 2011–2023.

Aurora, 1999. Availability of finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation. Document AU/217/99. Nokia.

Aurora, 2000. Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: description and baseline results. Document AU/271/00. UPC.

Aurora, 2001. Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation. Document AU/273/00. Texas Instruments.

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004

Aurora, 2001. Danish SpeechDat-Car digits database for ETSI STQ-Aurora advanced DSR. Document AU/378/01. Aalborg University.

Bagchi, D., Mandel, M.I., Wang, Z., He, Y., Plummer, A.R., Foslerlussier, E., 2015. Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Barfuss, H., Huemmer, C., Schwarz, A., Kellermann, W., 2015. Robust coherence-based spectral enhancement for distant speech recognition. arXiv:1604.03393v2

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third chime speech separation and recognition challenge: dataset, task and baselines. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The pascal chime speech separation and recognition challenge. Comput. Speech Lang. 27 (3), 621–633.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE 57 (8), 1408–1418.

Chen, J., Wang, Y., Wang, D.L., 2014. A feature study for classification-based speech separation at very low signal-to-noise ratio. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7089–7093.

Chen, S.F., Goodman, J., 1996. An empirical study of smoothing techniques for language modeling. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pp. 310–318.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Frontend factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Du, J., Wang, Q., Gao, T., Dai, L.-R., Lee, C.-H., 2014. Robust speech recognition with speech enhanced deep neural networks. In: Proceedings of Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 616–620.

Du, J., Wang, Q., Tu, Y., Bao, X., Dai, L., Lee, C., 2015. An information fusion approach to recognizing microphone array speech in the CHiME-3 challenge based on a deep learning framework. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (Rover). In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 347–352.

Gannot, S., Cohen, I., 2004. Speech enhancement based on the general transfer function GSC and postfiltering. IEEE Trans. Audio Speech Lang. Process. 12 (6), 561–571.

Garofalo, J., Graff, D., Paul, D., Pallett, D., 2007. CSRI (WSJ0) Complete LDC93S6A. Linguistic Data Consortium.

Ghahremani, P., BabaAli, B., Povey, D., 2014. A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2513–2517.

Glembek, O., Burget, L., Matejka, P., Karafiat, M., Kenny, P., 2011. Simplification and optimization of i-vector extraction. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4516–4519.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Commun. 16 (3), 261–291.

Graves, A., 2012. Supervised sequence labelling with recurrent neural networks. (Ph.D. thesis) University of Toronto.

Graves, A., Mohamed, A.-R., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649.

Griffiths, L.J., Jim, C.W., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. 30 (1), 27–34.

Heymann, J., Drude, L., Chinaev, A., Haebumbach, R., 2015. BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU).

Hirsch, H.G., 2002. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0. Technical Report. ETSI STQ-Aurora DSR Working Group.

Hori, T., Chen, Z., Erdogan, H., Hershey, J.R., Roux, J.L., Mitra, V., Watanabe, S., 2015. The MERL/SRI system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Jalalvand, S., Falavigna, D., Matassoni, M., Svaizer, P., Omologo, M., 2015. Boosted acoustic model learning and hypotheses rescoring on the CHiME3 task. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Jones, D.L., Ratnam, R., 2009. Blind location and separation of callers in a natural chorus using a microphone array. J. Acoust. Soc. Am. 126 (2), 895–910.

Keyi, A.E., Kirubarajan, T., Gershman, A., 2005. Robust adaptive beamforming based on the Kalman filter. IEEE Trans. Signal Process. 53 (8), 3032–3041.

Kneser, R., Ney, H., 1995. Improved backing-off for Mgram language modeling. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 181–184.

Li, B., Sim, K.C., 2013. Improving robustness of deep neural networks via spectral masking for automatic speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 279–284.

Li, J., Deng, L., Gong, Y., Haebumbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 22 (4), 745–777.

Li, W., Wang, L., Zhou, Y., Dines, J., Magimai-Doss, M., Bourlard, H., Liao, Q., 2014. Feature mapping of multiple beamformed sources for robust overlapping speech recognition using a microphone array. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 2244–2255.

Liu, Y., Zhang, P., Hain, T., 2014. Using neural network front-ends on far field multiple microphones based speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5542–5546.

Loesch, B., Yang, B., 2010. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions. In: Proceedings of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), pp. 41–48.

Ma, N., Marxer, R., Barker, J., Brown, G.J., 2015. Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Mcaulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. 34 (4), 744–754.

Mestre, X., Lagunas, M., 2003. On diagonal loading for minimum variance beamformers. In: Proceeding of International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 459–462.

Metze, F., Sheikh, Z., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q.B., Nguyen, V.H., 2013. Models of tone for tonal and non-tonal languages. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S., 2010. Recurrent neural network based language model. In: Proceedings of Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 1045–1048.

Mousa, A.E., Marchi, E., Schuller, B., 2015. The ICSTM+TUM+UP approach to the 3rd CHiME challenge: Single-channel LSTM speech enhancement with multi-channel correlation shaping dereverberation and LSTM language models. arXiv:1510.00268v1

Pang, Z., Zhu, F., 2015. Noise-Robust ASR for the third 'CHiME' Challenge Exploiting Time-Frequency Masking based Multi-Channel Speech Enhancement and Recurrent Neural Network. arXiv:1509.07211v1

Pearce, D., Hirsch, H., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 181–188.

Prudnikov, A., Korenevsky, M., Aleinik, S., 2015. Adaptive beamforming and adaptive training of DNN acoustic models for enhanced multichannel noisy speech recognition. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Renals, S., Swietojanski, P., 2014. Neural networks for distant speech recognition. In: Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA).

Sainath, T.N., Weiss, R.J., Wilson, K.W., Narayanan, A., Bacchiani, M., 2016. Factored Spatial and Spectral Multichannel Raw Waveform CLDNNs. Proc. ICASSP.

Sak, H., Senior, A., Beaufays, F., 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 338–342.

Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., Ma, M., 2014. Sequence discriminative distributed training of long short-term memory recurrent neural networks. In: Proceedings of Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 1209–1213.

Saon, G., Soltau, H., Nahamoon, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 55–59.

Sivasankaran, S., Nugraha, A.A., Vincent, E., Moralescordovilla, J.A., Dalmia, S., Illina, I., Liutkus, A., 2015. Robust ASR using neural network based speech enhancement and feature simulation. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Tachioka, Y., Watanabe, S., Le Roux, J., Hershey, J.R., 2013. Discriminative methods for noise robust speech recognition: a chime challenge benchmark. In: Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments, pp. 19–24.

Talmon, R., Cohen, I., Gannot, S., 2009. Relative transfer function identification using convolutive transfer function approximation. IEEE Trans. Audio Speech Lang. Process. 17 (4), 546–555.

Tu, Y.-H., Du, J., Dai, L.-R., Lee, C.-H., 2015. Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 61–65.

Veen, B.D., Buckley, K.M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE Signal Process. Mag. 10 (3), 4–24.

Vesel, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 2345–2349.

Vincent, E., Barker, J., Watanabe, S., Roux, J.L., Nesta, F., Matassoni, M., 2013. The second chime speech separation and recognition challenge: datasets, tasks and baselines. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Vincent, E., Gribonval, R., Plumbley, M., 2007. Oracle estimators for the benchmarking of source separation algorithms. Signal Process. 87 (8), 1933–1950.

Weng, C., Yu, D., Watanabe, S., Juang, B.-W., 2014. Recurrent deep neural networks for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5569–5572.

Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1), 7–19.

Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W.J., Espi, M., Higuchi, T., Araki, S., Nakatani, T., 2015. The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Zhang, C., Florencio, D., Ba, D.E., Zhang, Z., 2008. Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Trans. Signal Process. 10 (3), 538–548.

Zhao, S., Jones, D.L., Khoo, S., Man, Z., 2014. Frequency-domain beamformers using conjugate gradient techniques for speech enhancement. Journal of the Acoustical Society of America. 136 (3), 1160–1175.

Zhao, S., Xiao, X., Zhang, Z., Nguyen, T.N.T., Zhong, X., Ren, B., Wang, L., Jones, D.L., Chng, E.S., Li, H., 2015. Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Please cite this article as: Y. Tu et al., An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.12.004