INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On Design of Robust Deep Models for CHiME-4 Multi-Channel Speech Recognition with Multiple Configurations of Array Microphones

Yan-Hui Tu[1], *Jun Du*[1], *Lei Sun*[1], *Feng Ma*[2], and *Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]iFlytek Research, Hefei, Anhui, China
[3]Georgia Institute of Technology, Atlanta, Georgia, USA

tuyanhui@mail.ustc.ecu.cn, jundu@ustc.edu.cn, sunlei17@mail.ustc.ecu.cn,
fengma@iflytek.cn, chinhui.lee@ece.gatech.edu

Abstract

We design a novel deep learning framework for multi-channel speech recognition in two aspects. First, for the front-end, an iterative mask estimation (IME) approach based on deep learning is presented to improve the beamforming approach based on the conventional complex Gaussian mixture model (CGMM). Second, for the back-end, deep convolutional neural networks (DCNNs), with augmentation of both noisy and beamformed training data, are adopted for acoustic modeling while the forward and backward long short-term memory recurrent neural networks (LSTM-RNNs) are used for language modeling. The proposed framework can be quite effective to multi-channel speech recognition with random combinations of fixed microphones. Testing on the CHiME-4 Challenge speech recognition task with a single set of acoustic and language models, our approach achieves the best performance of all three tracks (1-channel, 2-channel, and 6-channel) among submitted systems.
Index Terms: CHiME challenge, deep learning, mask estimation, microphone array, robust speech recognition

1. Introduction

Automatic speech recognition (ASR) in distant-talking scenarios based on the use of microphone arrays has become an important part of everyday life with the emergence of speech-enabled applications on multi-microphone portable devices due to its convenience and flexibility. However, the speech signals recorded by distant microphones are often corrupted by reverberation and background noise, leading to considerable degradation in ASR performance, particularly at low signal-to-noise ratios (SNRs). Speech enhancement algorithms that reduce noise without much damaging the target speech are therefore desired for improving the ASR performance and robustness. For multi-channel speech enhancement, representative algorithms in this category include multi-channel Wiener filtering [1], blind source separation [2], and beamforming [3, 4]. And beamforming is a popular approach in the CHiME-3 Challenge [5], which focuses on solving ASR problems in real-world applications. For example, the minimum variance distortionless response (MVDR) beamformer was used extensively in a few top-performing CHiME-3 ASR systems [6, 7]. A key to achieving a high-quality beamformer is how to construct a steering vector that represents the acoustic propagation [8]. Conventionally, some *a priori* knowledge is used to construct the steering vector, e.g., the geometry of the microphone array and the direction of arrival (DOA) information. But its robustness often becomes a problem in real-life environments where the acoustic propagation information is not known and difficult to estimate accurately. In [4], a method was developed to steer a beamformer using the time-frequency (T-F) masks estimated by a complex Gaussian mixture model (CGMM), which was demonstrated to be beneficial to the top-performing CHiME-4 ASR systems [9, 10] as well.

Deep learning techniques are becoming increasingly popular in many speech research areas, notably ASR [11]. In [12, 13], deep neural networks (DNNs) were utilized for single-channel enhancement and shown to be superior to some early speech enhancement algorithms in improving some objective measures, such as short-time objective intelligibility (STOI) [14] and segmental SNR (SSNR, in dB) [15]. Different neural network architectures have been adopted in single-channel speech enhancement for ASR, and they have demonstrated a significant increase in ASR performance [16, 17, 18]. The input features of these approaches are magnitude or log-magnitude spectra in the short-time-Fourier-transform (STFT) domain [19]. The ideal ratio mask (IRM) [20] has also been shown to obtain a good speech enhancement performance.

In this paper, we propose to improve multi-channel speech recognition via a deep learning framework. First, a closed-loop approach to beamforming by leveraging upon information obtained via iterative neural network based IRM estimation and ASR based voice activity detection (VAD) [21] has been proposed as our front-end system, which preprocesses random channels data to single channel beamfomed data. Due to the introduction of data-driven approach for multi-channel enhancement, the proposed approach is robust to the space geometry relation of microphone array. Meanwhile, a powerful back-end system is also designed for improving the recognition performance. We first explore a key technique used in building our CHiME-3 ASR system [22], namely data augmentation, in the spirit of fusing front-end features of both processed and unprocessed multi-channel data. Next, the acoustic model in the framework of hidden Markov model (HMM) is upgraded via deep convolutional neural networks (DCNNs) [23] with more layers and a smaller filter size than conventional convolutional neural networks (CNNs) [24]. Finally, the language models (LMs) based on long short-term memory (LSTM) [25] are used with a combination of forward and backward LSTMs to further improve the ASR performance. Testing on the CHiME-4 three tracks (1-channel, 2-channel, and 6-channel microphone array data), our proposed framework achieves the best performance among all submitted systems. This work is an extension of the recently disclosed version [26], with more experiments on 2-channel and 1-channel cases to demonstrates its robustness to the random combination of fixed microphones.

http://dx.doi.org/10.21437/Interspeech.2017-853

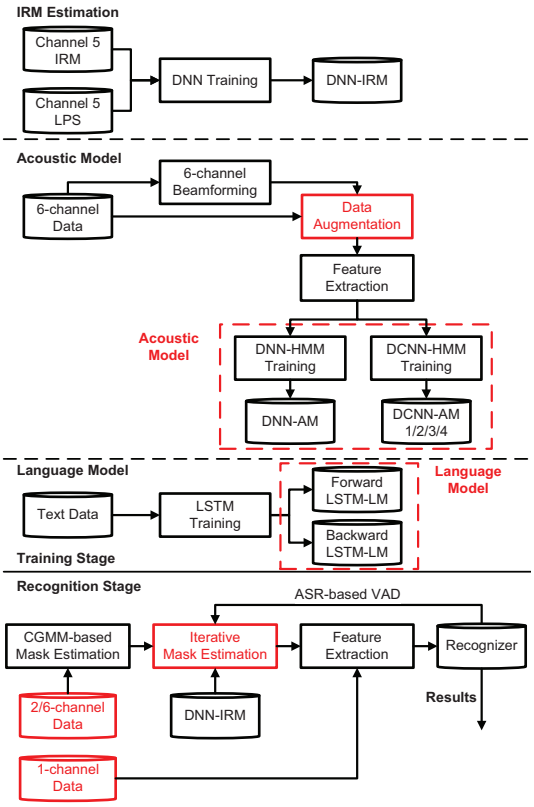

Figure 1: *A block diagram of the proposed deep learning framework, which consists of front-end beamforming and back-end acoustic/language models.*

2. Proposed Deep Learning Framework

A block diagram of the proposed deep learning framework is shown in Figure 1. The training stage consists of three parts, i.e., IRM estimation, acoustic model and language model. For the IRM estimation, the DNN-IRM model is trained using the log-power spectral (LPS) features of data from channel 5 (the main channel of microphone array in CHiME-4 task) as input features and the IRM as output target. The LPS features offering perceptually relevant parameters are adopted [15]. IRM is used to represent the speech presence probability at each time-frequency point in speech separation [27]. Then, the acoustic models, including one DNN-HMM and four DCNN-HMMs, are trained using beamformed data as well as data from channels 1, 3, 4, 5, and 6. Only the input features of DNN are concatenated from log mel-filterbank (LMFB) and feature-space maximum likelihood regression (fMLLR) features [28]. The difference of the DCNN1-4 is the filter size (3×5 and 3×3) and input features. Finally, the forward and backward LSTM-based language models are also constructed.

The recognition process of 2-channel and 6-channel data is divided into four successive steps, namely, beamforming initialization, DNN-based signal statistics (IRM and ASR-based VAD) estimation, beamforming, and recognition. First, beamformed speech from the multi-channel input is initialized and a time-frequency (T-F) mask of test speech is obtained by CGMM-based beamforming. Then, the IRM estimated by a well-trained DNN-IRM model and the ASR-based VAD information are used to improve the initial mask. Next, the improved mask is adopted to steer the beamformer, thereby obtaining the beamformed speech for ASR. Finally, multiple acoustic models are first fused at the state level, and then first-pass decoding is performed with 3-gram to generate lattices as the hypotheses, which are subsequently served for the second-pass decoding with LSTM-based LMs. For the 1-channel data, the original data is directly fed into recognizer. The details are presented in the following subsections and the readers can also refer to [26].

3. Iterative Mask Estimation based Beamforming

We use minimum variance distortionless response (MVDR) beamformer which maximizes the signal-to-noise ratio (SNR) of the beamformer output in each frequency bin k, leading to the beamformer coefficients [29]:

$$\boldsymbol{w}(k) = \frac{\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}{\boldsymbol{g}^{H}(k)\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}, \tag{1}$$

where $\boldsymbol{g}(k)$ is the signal propagation vector, which is in the same form as the so-called steering vector in the literature of array beamforming [8]; $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ and $\boldsymbol{R}_{\boldsymbol{nn}}(k)$ are the spatial correlation matrix of target and noise, respectively.

The spatial correlation matrix can be estimated by the time-frequency mask $M(k,l)$, which represents the probability of the T-F unit (k,l) containing the target speech signal. The key to this approach is unsupervised and accurate estimation of the spectral masks that indicate the presence and absence of speech T-F units. In [4], the CGMM-based approach to estimating the masks has been proposed.

3.1. The proposed iterative mask estimation procedure

In this section, we discuss the iterative mask estimation (IME) with DNN-based IRM and ASR-based VAD to improve the masks estimated by the CGMM-based approach. The procedure of IME is described as follows:

Step 1: Estimate the initial mask for each T-F unit (k,l), denoted as $M_{\text{CGMM}}(k,l)$, using the CGMM-based approach.

Step 2: Steer the beamformer with the estimated mask and obtain the beamformed speech.

Step 3: Feed the DNN-IRM model with the beamformed speech from Step 2 to estimate IRM, denoted as $M_{\text{DNN}}(k,l)$.

Step 4: Perform the first-pass decoding with the beamformed speech from Step 2 to get the ASR-based VAD, denoted as $M_{\text{ASR}}(k,l)$.

Step 5: Combine $M_{\text{CGMM}}(k,l)$ in Step 1 with $M_{\text{DNN}}(k,l)$ in Step 3 or/and $M_{\text{ASR}}(k,l)$ in Step 4 to generate the improved mask.

Repeat Steps 2–5 for N iterations.

3.2. Improving mask estimation by DNN-based IRM

We use a DNN-IRM to predict the mask, $M_{\text{DNN}}(k,l)$, representing the speech presence probability at every T-F unit given the input LPS features of enhanced speech obtained at Step 2 in Section 3.1. And the estimated $M_{\text{DNN}}(k,l)$ is combined with $M_{\text{CGMM}}(k,l)$ to yield an improved mask $M_1(k,l)$, i.e.,

$$M_1(k,l) = \sqrt{M_{\text{CGMM}}(k,l)M_{\text{DNN}}(k,l)}. \tag{2}$$

This process can repeat iteratively following the Steps 2-5 in Section 3.1.

3.3. Improving mask estimation by ASR-based VAD

The VAD information, $M_{\text{ASR}}(k,l)$, from the segmentation results of the speech recognizer using beamformed speech at each frame is used to further improve the mask estimation as

$$M_2(k,l) = M_{\text{CGMM}}(k,l) M_{\text{ASR}}(k,l). \quad (3)$$

Please note that Eq. (3) only uses the ASR-based VAD information to improve the CGMM-based mask. According to the Step 5 in Section 3.1, if both DNN-based mask and ASR-based mask are adopted, $M_{\text{CGMM}}(k,l)$ in Eq. (3) should be replaced by $M_1(k,l)$ in Eq. (2), i.e.,

$$M_3(k,l) = M_1(k,l) M_{\text{ASR}}(k,l). \quad (4)$$

Similar to $M_1(k,l)$, $M_2(k,l)$ and $M_3(k,l)$ can be iteratively refined by repeating Steps 2-5 of Section 3.1.

4. Back-End Recognizer Design

4.1. DCNN-HMM based acoustic modeling

In our CHiME-3 system [28], the LSTM-HMM was adopted and combined with DNN-HMM. In CHiME-4, we use DCNN-HMM, which is found more effective. The model consists of the input layer, 4 blocks with different sizes of feature maps, one fully connected (FC) layer and the softmax output layer. For each block, there are six nets and a max-pooling layer. Each net is a nonlinear transformation, and the input feature map is processed with a conventional convolution ReLU layer with batch normalization. We design 3×3 and 3×5 kernels to build two DCNN models by considering the complementarity of different kernel sizes. Moreover, it is important for speech data to use a small kernel in DCNN due to the low resolution of acoustic features. Unlike the DNN-HMM where LMFB and fMLLR features are concatenated, these two feature types are separately used to generate two sets of DCNN-HMMs. Finally, four DCNN-HMMs are built, which are combined with the DNN-HMM using the state-level model fusion [28] for better recognition accuracy.

4.2. LSTM-based language modeling

The language model plays an important role in ASR, which incorporates syntactical and semantical constraints in the decoding process. A powerful language model can help significantly improve ASR performance. In this work, we use the LSTM-based language model [30, 25] rather than the simple recurrent neural network (SRNN) based language model officially provided [31]. First, the input word sequence is represented by the one-hot encoding as a 4985-dimensional vector, which may result in data sparsity problem. Then, word embedding is adopted to provide a better representation to alleviate the problem of data sparsity. One LSTM layer with 1024 memory cells is followed by a 512-dimensional projection layer which can be interpreted as projecting the input words to a continuous space. The last two layers are the fully connected layer with 1024 nodes and the output layer with the vocabulary size of 4985. To fully utilize the directional information of word sequences, two LSTMs to model the text line from different directions, namely forward LSTM and backward LSTM, are designed. It should be emphasized that the combination of these two LSTM-based LMs generates much better recognition results than just using one single bi-directional LSTM-based LM.

5. Experimental Evaluation

We present the experimental evaluation of our framework in the CHiME-4 task [31], which was designed to study real-world ASR scenarios where a person is talking to a mobile tablet device equipped with 6 microphones in a variety of adverse environments. CHiME-4 offers three tasks (1-channel/1ch, 2-channel/2ch, and 6-channel/6ch) with different testing scenarios. More detailed information of CHiME-4 can refer to [5, 31].

5.1. Experiments on iterative mask estimation

In this subsection, the baseline ASR recognition system officially provided [31] is used to evaluate the different beamformers on the test sets of real data. The acoustic model is DNN-HMM discriminately trained with state-level minimum Bayes risk (sMBR) criterion. The input of DNN-HMM is a 440-dimensional feature vector extracted from channel 5, consisting of 40-dimensional fMLLR with an 11-frame expansion. The language models are 5-gram with Kneser-Ney (KN) smoothing for the first-pass decoding and the SRNN-based language model for rescoring. The DNN-IRM is trained using 7 frames of 257-dimension LPS features of channel 5. The DNN-IRM architecture is 1799-2048*3-257, namely 257*7 dimension for LPS input features, 3 hidden layers with 2048 nodes for each, and 257 nodes for the output T-F IRM. For the DNN-IRM fine-tuning, the learning rate is set to 0.01 for 50 epochs, and the mini-batch size is 128. In the training stage, only the simulation data are adopted with the input/output pairs of channel 5 speech and the corresponding IRMs. For the CGMM-based beamforming, the multi-channel STFT coefficients are extracted from the test speech at a 16 kHz sampling frequency using a Hanning window of length 512 and shift of 256, resulting in 257 frequency bins.

Table 1 presents the word error rate (WER) comparison among the CGMM-based beamformer and its improved versions by incorporating DNN-based IRM and ASR-based VAD for the 2ch and 6ch tracks on the test sets of real data. "+DNN-IRM" and "+ASR-VAD" denote the iterative mask estimation in the first iteration via Eq. (2) and the second iteration via Eq. (4), respectively. First, the DNN-IRM based approach achieves consistent and significant improvements of recognition performance over the CGMM-based method, yielding average relative WER reductions of 12.2% and 20.7% across all test sets for 2ch and 6ch tracks, respectively. The performance gain of 6ch track is more significant than that of 2ch for the DNN-IRM based approach because the IRM estimation of 6ch is more accurate than that of 2ch. Second, the ASR-based VAD in the second iteration for iterative mask estimation achieved significant and stable recognition performance gains across all test sets of real data, with average relative WER reductions of 5.8% and 9.2% over DNN-IRM in the first iteration for 2ch and 6ch tracks, which demonstrates its strong complementarity with both CGMM-based and DNN-based mask estimation. Overall, the proposed IME approach generates the relative WER reductions of 17.3% and 28.0% over the CGMM-based approach for 2ch and 6ch tracks across all test sets, respectively, which demonstrates its effectiveness and robustness to multi-channel speech recognition with a random combination of fixed microphones.

Table 1: *WER comparison of different beamformers on the test sets of real data for 1ch, 2ch and 6ch tracks using the official DNN-HMM acoustic model.*

Beamformer	Track	BUS	CAF	PED	STR	AVG
No	1ch	36.17	24.86	18.93	13.93	23.47
CGMM	2ch	20.08	12.85	9.68	9.90	13.13
	6ch	13.24	8.12	6.67	6.03	8.54
+DNN-IRM	2ch	15.76	11.30	9.06	10.01	11.53
	6ch	9.63	5.98	5.85	5.62	6.77
+ASR-VAD	2ch	14.88	10.74	8.58	9.25	10.86
	6ch	7.80	5.75	5.42	5.64	6.15

Table 2: *The settings of different DCNN-HMMs.*

Acoustic Model	Input Feature	Kernel Size
DCNN1-HMM	LMFB	3×3
DCNN2-HMM	LMFB	3×5
DCNN3-HMM	fMLLR	3×3
DCNN4-HMM	fMLLR	3×5

5.2. Experiments on acoustic and language models

5.2.1. Training data augmentation

Different from the DNN-HMM in Section 5.1 where only fMLLR features were used, the input feature vector for data augmentation experiments consists of 42-dimensional LMFB, 40-dimensional fMLLR, and 20-dimensional i-vector [22]. For both LMFB and fMLLR, the first-order and second-order derivatives with 9-frame expansion are adopted, yielding a 2234-dimensional (2234=42*3*9+40*3*9+20) feature vector fed to the input layer of DNN. We use 7 hidden layers with 2048 nodes for each layer and 1965 nodes for the output layer. Other configurations follow the Kaldi setup officially provided [31]. And the test data of 2ch and 6ch tracks are beamformed by the proposed IME approach. The first row of Table 3 lists WER results of data augmentation for DNN-HMM acoustic model on the test sets of real data for three tracks. By comparison with the results in Table 1 (the first row for 1ch, the last two rows for 2ch and 6ch), it is clear that adding all channels of the original noisy speech except channel 2 and beamformed speech as augmented training data remarkably reduces the WERs over the baseline system only trained with channel 5 data, with average relative reductions of 37.9%, 34.3% and 29.6% for 1ch, 2ch and 6ch tracks across all test sets, respectively.

5.2.2. The ensemble of DNN-HMM and DCNN-HMMs

For DCNN-HMMs, four models are built with different settings of input features, kernel sizes as shown in Table 2. The learning rate of DCNN training is set to 0.002, and the batch size is 2048. Batch normalization was used to accelerate the training. The "SRNN" LM block of Table 3 shows the WER comparison with different acoustic models on the test sets of real data for the three tracks. First, DCNN1-HMM with LMFB as input features outperforms DNN-HMM with the concatenation of LMFB and fMLLR features, e.g., with an average relative WER reduction of 10.6% for the 6ch track, demonstrating the importance of deeper architectures with convolutional layers. Second the ensemble of all five models (one DNN-HMM and four DCNN-HMMs) yields the relative WER reductions of 23.7%, 24.4%, 25.6% over the DNN-HMM system for for 1ch, 2ch and 6ch tracks, respectively.

Table 3: *WER comparison of different acoustic models and language models on the test sets of real data for 1ch, 2ch and 6ch tracks.*

LM	AM	Track	BUS	CAF	PED	STR	AVG
SRNN	DNN	1ch	19.61	16.16	12.95	9.62	14.58
		2ch	9.57	6.54	5.81	6.63	7.14
		6ch	5.20	3.98	3.48	4.69	4.33
	DCNN1	1ch	22.56	16.75	12.80	9.79	15.47
		2ch	9.44	6.89	5.44	5.72	6.87
		6ch	4.60	3.64	3.49	3.74	3.87
	Ensemble	1ch	16.56	11.26	9.04	7.60	11.12
		2ch	7.27	5.23	4.43	4.69	5.40
		6ch	3.96	2.95	2.82	3.16	3.22
LSTM	Ensemble	1ch	14.10	9.64	6.89	5.98	9.15
		2ch	5.16	3.83	3.18	3.49	3.91
		6ch	2.65	2.09	1.74	2.48	2.24

5.2.3. LSTM-based LMs

The "LSTM" LM block of Table 3 presents the WERs of combining forward and backward LSTM-based language models on the test sets of real data for the three tracks. "SRNN" denotes the simple RNN-based language model officially provided. From "SRNN" to "LSTM", average relative WER reductions of 17.7%, 27.6% and 30.4% are achieved for 1ch, 2ch and 6ch tracks, respectively, which implies LSTM-based LMs can be more powerful with the better beamformer and acoustic modeling. Overall, by the integration of the proposed beamforming, data augmentation, acoustic and language modeling, the deep learning framework can achieve the average WERs of 9.15%, 3.91% and 2.24% for 1ch, 2ch and 6ch tracks on the test sets of real data, respectively, which are the best results among all submitted systems in CHiME-4 Challenge. Based on the above results of all three tracks, we conclude that the design of deep models in this study for both front-end and back-end could be quite effective and robust to multi-channel speech recognition with a random combination of fixed microphones or even a single-microphone speech recognition in the testing stage.

6. Conclusions

In this paper, we improve multi-channel speech recognition with a random combination of fixed microphones via a deep learning framework. By integrating five key techniques, i.e., (i) iterative mask estimation (IME) based beamforming, (ii) data augmentation with both processed and unprocessed speech, (iii) detailed acoustic modeling using multiple DNN-based and DCNN-based acoustic models, (iv) detailed language modeling using both forward and backward LSTMs, and (v) system combination , our system has achieved the lowest WERs among all participating systems for three tracks of ASR performance evaluation (1-channel, 2-channel and 6-channel) on test sets of real data in the recent CHiME-4 Challenge.

7. Acknowledgment

This work was partially funded by The National Key Research and Development Program of China (Grant No.2016YFB1001300), National Natural Science Foundation of China grant no U1613211, National Natural Science Foundation of China Grant No 61671422. This work was also supported by Samsung.

8. References

[1] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[3] A. Krueger, E. Warsitz, and R. Haebumbach, "Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, 2011.

[4] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[6] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. F. C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[7] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. L. Roux, V. Mitra, and S. Watanabe, "The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[8] B. D. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 4–24, 1988.

[9] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The ustc-iflytek system for chime-4 challenge," *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, 2016.

[10] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schlter, H. Ney, R. Haeb-Umbach, and A. Mouchtaris, "The rwth/upb/forth system combination for the 4th chime challenge evaluation," *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments (CHiME 2016)*, 2016.

[11] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. W. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, p. 82, 2012.

[12] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[13] Y. Tu, J. Du, Y. Xu, L. Dai, and C. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *International Symposium on Chinese Spoken Language Processing.(ISCSLP)*, 2014.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timecfrequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2008.

[16] J. Du, Y. Tu, L. Dai, and C. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[17] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[18] Y. Tu, J. Du, L. Dai, and C. Lee, "Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, 2015.

[19] R. J. Mcaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[20] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[21] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[22] Y.-H. Tu, J. Du, Q. Wang, X. Bao, L.-R. Dai, and C.-H. Lee, "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 and Language*, 2016.

[23] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2016.

[24] O. Abdelhamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, 2012, pp. 4277–4280.

[25] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *CoRR*, vol. abs/1602.02410, 2016. [Online]. Available: http://arxiv.org/abs/1602.02410

[26] Y.-H. Tu, J. Du, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Submitted to IEEE/ACM Transactions on Audio Speech and Language Processing*.

[27] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," *Blind Source Separation*, pp. 349–368, 2014.

[28] J. Du, Q. Wang, Y.-H. Tu, X. Bao, L.-R. Dai, and C.-H. Lee, "An information fusion approach to recognizing microphone array speech in the chime-3 challenge based on a deep learning framework," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[29] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[30] M. Sundermeyer, R. Schluter, and H. Ney, "Lstm neural networks for language modeling," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2012.

[31] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, 2016.