ACOUSTIC MODELING FOR DISTANT MULTI-TALKER SPEECH RECOGNITION WITH SINGLE- AND MULTI-CHANNEL BRANCHES

Naoyuki Kanda[1]*, Yusuke Fujita*[1]*, Shota Horiguchi*[1]*, Rintaro Ikeshita*[1]*, Kenji Nagamatsu*[1]*, Shinji Watanabe*[2]

[1]Hitachi Ltd., Japan

{naoyuki.kanda.kn, yusuke.fujita.su, shota.horiguchi.wk, rintaro.ikeshita.bk, kenji.nagamatsu.dm}@hitachi.com

[2]Johns Hopkins University, USA

shinjiw@ieee.org

ABSTRACT

This paper presents a novel heterogeneous-input multi-channel acoustic model (AM) that has both single-channel and multi-channel input branches. In our proposed training pipeline, a single-channel AM is trained first, then a multi-channel AM is trained starting from the single-channel AM with a randomly initialized multi-channel input branch. Our model uniquely uses the power of a complemental speech enhancement (SE) module while exploiting the power of jointly trained AM and SE architecture. Our method was the foundation for the Hitachi/JHU CHiME-5 system that achieved the second-best result in the CHiME-5 competition, and this paper details various investigation results that we were not able to present during the competition period. We also evaluated and reconfirmed our method's effectiveness with the AMI Meeting Corpus. Our AM achieved a 30.12% word error rate (WER) for the development set and a 32.33% WER for the evaluation set for the AMI Corpus, both of which are the best results ever reported to the best of our knowledge.

Index Terms— Acoustic model, speech recognition, speech enhancement, deep learning

1. INTRODUCTION

Thanks to the recent advances in deep-learning-based speech recognition [1–3], word error rates (WERs) for some datasets have become close to (e.g., Switchboard [4, 5]) or just below (e.g., LibriSpeech in [6] and [7]) the level of human transcribers. However, despite this progress, noise and reverberation still severely degrade WERs. Multi-talker speech recognition with a distant microphone is one of the most difficult settings for speech recognition because of the difficulty of separating the target speaker's speech from other speech. One example is meeting speech recognition with a distant microphone. Another example is conversational speech recognition in a home environment, which helps develop a highly intelligent personal agent. It is known that the WERs for such situations are still very high (30% [8] to 80% [9]) even with state-of-the-art speech recognizers.

One way to improve distant multi-talker speech recognition is to improve the robustness of acoustic models (AMs) with data augmentation techniques [10], better training objectives [7, 11–14], improved model architecture [15, 16], etc. Another way is improving speech enhancement (SE) techniques. Recent progress in deep-learning-based speech separation showed dramatic improvement of WERs in many scenarios [17–23]. However, the 5th CHiME Speech Separation and Recognition Challenge (CHiME-5) [9] revealed to us that conventional speech separation techniques led only marginal improvements under noisy far-field multi-talker conditions. We internally evaluated deep clustering [19], permutation invariant training [20], chimera++ networks [21], and a speaker-aware neural beamformer [22, 23], but none of them successfully improved the accuracy for the CHiME-5 dataset. We eventually found that the parameter-based speaker adaptation of the neural beamformer slightly improved the WER [24], but the improvement by that method remained minor compared with the improvements reported in past literature under easier conditions (e.g., [17, 18]).

Recently, the unified optimization of AM and SE was proven effective for noisy conditions [15, 25–30]. For example, some papers [27, 28] proposed a convolutional neural network (CNN)-based multi-channel SE block that was jointly trained with a neural network-based AM. In [29], a speaker-aware SE block was jointly trained with an AM by using an additional speaker representation network. Paper [30] proposed the combination of a trainable multi-channel beamformer with end-to-end speech recognition. All the works were promising, and our work is following in the same direction as previous studies on the joint training of the AM and SE.

In our paper, we propose novel heterogeneous-input multi-channel acoustic modeling. The unique feature of our method is that our model architecture has two input branches: one for single-channel signals and another for multi-channel signals. We propose an AM training scheme where we first train a single-channel AM, then a multi-channel AM is trained starting from the single-channel AM with a randomly initialized multi-channel input branch. Our experiments in CHiME-5 and AMI Corpus showed that this architecture and training scheme enables us to use the power of a complemental SE module while exploiting the power of jointly trained AM and SE architecture.

In summary, our proposed AM has the following advantages:

- It provides an easy way to use a complemental SE module while exploiting the power of a jointly trained AM and SE model.
- It produces state-of-the-art accuracy. It was the foundation for the second-best result at the CHiME-5 competition [9]. It also achieved, to the best of our knowledge, the best WER ever reported for AMI Corpus.

This paper corresponds with the extended investigation of our CHiME-5 paper [24]. We present systematic experimental analyses (including the AMI evaluation to further investigate the effectiveness of our proposed AM) that we could not conduct during the competition period due to the limited development time.

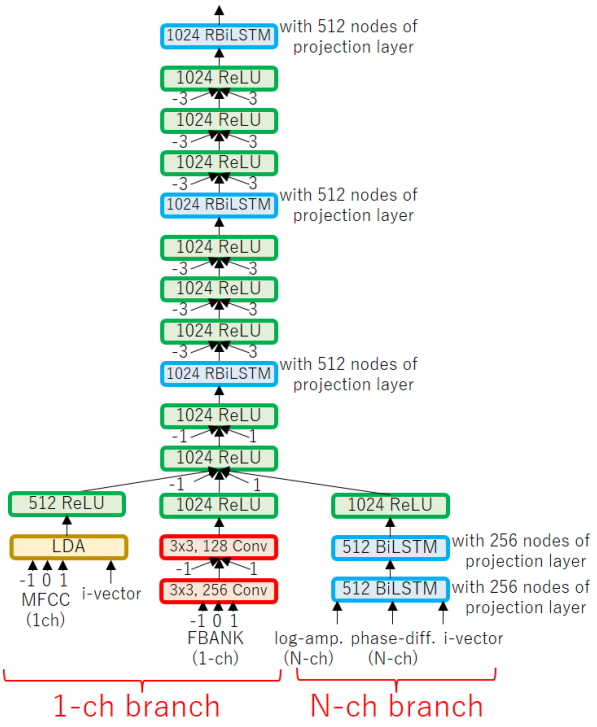

Fig. 1. *AM architecture. A number with an arrow indicates a time splicing index, which forms the basis of TDNN [31].*

2. HETEROGENEOUS-INPUT MULTI-CHANNEL AM

2.1. Overview of AM architecture

Figure 1 depicts the multi-channel AM architecture that we used for our evaluation (both for CHiME-5 and AMI). In this figure, the red blocks represent a convolutional neural network, the blue blocks represent a bidirectional long short-term memory (BiLSTM) or a residual BiLSTM (RBiLSTM) that we proposed in [24], and the green blocks represent a time-delay neural network [31].

The unique part of this model architecture is in its input branch. This model has input branches for single-channel features and an input branch that accepts multi-channel features. We use mel-frequency cepstral coefficients (MFCCs) and log-Mel-filterbank (FBANK) as input for the single-channel branch. On the other hand, we use two types of features that represent multi-channel input signals for the multi-channel branch. One feature is log amplitude $\log|x_{i,f,t}|$ for each microphone i $(=1,...,N)$, time frame t, and frequency bin f, where N is the number of input channels. Another feature is the phase difference between each microphone and the first microphone as follows.

$$\cos(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2,...,N), \tag{1}$$

$$\sin(\angle(x_{i,f,t}) - \angle(x_{1,f,t})) \qquad (i = 2,...,N). \tag{2}$$

2.2. Training strategy for the proposed AM

Our proposed training pipeline for the heterogeneous-input multi-channel AM is as follows.

1. **1-ch AM training.** A neural network without N-ch input branch is initialized with random values and trained using 1-ch training data based on an ASR training objective (e.g., lattice-free maximum mutual information (LF-MMI) [13]) .
2. **N-ch AM training.** A randomly initialized N-ch input branch is added to the trained 1-ch AM, and parameters are trained using N-ch training data based on the ASR training objective.

This training pipeline is inspired by the curriculum learning concept [32] in which the neural network is first trained with an easy problem and then trained with more difficult problems.

One important point of our training scheme is in the training of the 1-ch AM (the first step of the training pipeline). If robust spectral representation of acoustics is learned in this step, the N-ch input branch can concentrate on learning the spacial beamforming property. To do so, we applied many data augmentation techniques [10, 24, 33–35] for the 1-ch AM training, such as speed and volume perturbation [34], reverberation, noise perturbation [35], and bandpass perturbation [24]. We also used both enhanced and non-enhanced speech signals for training, which were also effective in our experiment. Since the objective of the first step is learning robust spectral representation, we applied as many data augmentation methods as possible even if it is a phase-destructive method.

In addition, at the second step of the training pipeline above, we tested the two training options below.

Full update: Entire parameters are updated.

Partial update: Only parameters in the N-ch input branch are updated.

The partial update aims to mitigate the overfitting problem by reducing the number of training parameters. As the next section shows, this trick was very effective in our experiments.

3. EXPERIMENT

3.1. Experiments on CHiME-5

3.1.1. Experimental settings

We first evaluated our method with the CHiME-5 dataset [9]. This dataset contains home party recordings in which four participants spoke spontaneously in the kitchen, dining room, and living room. Six microphone arrays, each of which had four microphones, were used for the recordings. In addition, each participant put on a binaural microphone. While both microphone array data and binaural microphone data were allowed to be used for training AMs and LMs, only microphone array data was used for evaluation. There were two tasks; one used only reference array data (single array track) and one used all the arrays (multiple array track). In this paper, we used the setting for a single array track. A baseline program based on the Kaldi toolkit [36] and ESPnet [37] was released by the CHiME-5 organizers, and our experiments were conducted based on Kaldi. Due to very difficult recording conditions, the WER for the CHiME-5 official baseline system was 81.1%.

Training was 40.6 hours, development was 4.5 hours, and evaluation data was 5.2 hours. As explained in Section 2.2, we applied speed and volume perturbation (×3) [34], reverberation and noise perturbation (×2), [35] and bandpass perturbation (×2) [24] for the first step of the training pipeline, which produced roughly 4,500 hours of training data. Table 1 shows the effect these data augmentation techniques had using the CHiME-5 baseline TDNN architecture. As shown in the table, the more data augmentation techniques were applied, the better WER became. We used the official lexicon and language model for decoding. Note that we only reported

Table 1. *Effect of data augmentation for the CHiME-5 baseline TDNN [9]. This table is cited from Table 1 of [24].*

Data	Epochs	SP/VP	RP/NP	BP	Ref-Array
$W + R_1$	4	√			79.65
$W + R_1 + B_1$	4	√			78.72
$W + R_{1..6} + B_{1..6}$	4	√			78.51
$W + R_{1..6} + B_{1..6}$	2	√	√		77.26
$W + R_{1..6} + B_{1..6}$	1	√	√	√	**76.31**

SP/VP: speed & volume perturb. RP/NP: reverb. & noise perturb., BP: bandpass perturb.

In data column, W: worn mic., R_i: raw 1ch of i-th array, B_i: BeamformIt 1ch of i-th array.

Table 2. *WERs (%) of the single-channel / multi-channel AMs with different speech enhancement methods for CHiME-5.*

AM	Frontend for 1-ch input	Frontend for 4-ch input	Ref-Array
1-ch	Raw(CH1)	-	66.65
1-ch	WPE	-	66.20
1-ch	WPE + NN-MVDR	-	63.97
4-ch	Raw(CH1)	Raw	63.16
4-ch	WPE	WPE	62.74
4-ch	WPE + NN-MVDR	WPE	**61.91**

WERs of development data because transcriptions of the evaluation data were not published and we were not able to calculate the WER of the evaluation data.

Figure 1 shows our AM architecture. We used a 40-dim MFCC and a 40-dim FBANK without normalization for the 1-ch input branch. In addition, a 100-dim i-vector was extracted every 100 msec and it was used for online speaker/environment normalization [38]. For the N-ch input branch, we used a 257-dim log amplitude for each microphone and a 510-dim phase difference (255-dim for $\cos()$ and 255-dim for $\sin()$) for the $N-1$ microphone pairs. The utterance-based mean and variance normalization were applied for the log amplitude feature, while no normalization was applied for the phase difference feature. Through our AM training, we used LF-MMI as a training objective. We applied $l2$-regularization and cross-entropy-regularization [13] with scales of 0.00005 and 0.1, respectively. In addition, we used a backstitch technique [39] with the backstitch scale 1.0 and backstitch interval 4. Although we also applied lattice-free state-level minimum Bayes risk (LF-sMBR) training [7] in the CHiME-5 competition, we omitted it to simplify the experiments. Instead of LF-sMBR, we investigated a different number of training iterations of LF-MMI with different training schemes that we were not able to fully investigate during the CHiME-5 competition period. Section 3.1.3 discusses the results of this investigation.

3.1.2. Comparison of single-channel and multi-channel AMs

Table 2 compares WERs with a 1-ch AM that is trained for the first stage of our training pipeline and a 4-ch AM after adding the 4-ch input branch. We show results with a weighted prediction error (WPE) [40] and with a mask-estimation neural-network-based MVDR beamformer (NN-MVDR) with a speaker adaptation technique [24], which we found to be the best for the CHiME-5 dataset. Note that results with BeamformIt [41] were omitted as it was found to be ineffective for this dataset [24].

Table 2 shows that the combination of WPE and NN-MVDR produced about 3.7% of the absolute WER improvement and achieved 63.97% of the WER. Then, we found that the 4-ch AM achieved a better, 63.16% WER without using any other SE techniques.[1] By applying the WPE and NN-MVDR, we further obtained about a 1.3-point absolute WER improvement, and we finally achieved 61.91% of the WER. We noted that applying NN-MVDR achieved additional improvement for our 4-ch AM, which was realized through the 1-ch input branch. This is a very unique property of our AM that has a 1-channel input branch as well as a N-channel input branch.

[1]Note that the values in Table 2 are slightly better than the values reported in [24] due to correcting the misconfiguration of the decoder settings that we found after our CHiME-5 submission.

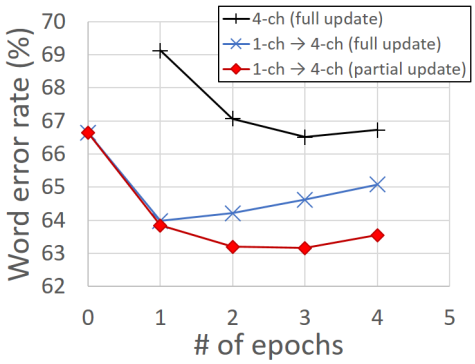

Fig. 2. *Comparison of AM training scheme.*

Table 3. *WERs (%) with different features for 4-ch input branch for CHiME-5.*

AM	Feature for 4-ch input branch		Ref-Array
	log-amp.	phase-diff.	
1-ch	n/a	n/a	66.65
4-ch		√	65.06
4-ch	√		63.23
4-ch	√	√	**63.16**

3.1.3. Comparison of training strategy

In our training pipeline, we first trained a single-channel AM and then continued the training with a N-channel input branch. We evaluated the effectiveness of this training procedure. Figure 2 compares the three training strategies. For the "4-ch (full update)" setting, we initialized entire parameters of 4-ch AM and trained all parameters based on LF-MMI. As shown in the figure, this method produced significantly worse results compared with the proposed two-pass training scheme ("1-ch → 4-ch (full update)"). By comparing "full update" and "partial update" (explained in Section 2.2), we found that the partial update effectively mitigated the overfitting problem and achieved the best results.

3.1.4. Comparison of multi-channel features

We also compared the input features for the N-ch input branch. We conducted an ablation study in which we used only the log-amplitude feature or phase-difference feature. Table 3 shows that both the log-amplitude feature and the phase-difference feature improved the accuracy while the improvement by the phase-difference feature was relatively small. Interestingly, we observed some improvement even when we used only the phase-difference feature for the 4-ch input branch. Note that this result was not obtained by the AM with only phase information; our AM always accepts spectral information via the 1-ch input branch. We thought that the phase information contained some complemental information for the spectral information, such as information for voice activity detection.

Table 4. *WERs (%) of the 1-ch / 8-ch AMs for AMI Corpus in the settings of far-field ASR.*

AM	Training data for 8ch-AM		Evaluation data		Dev	Eval
	1ch-branch	8ch-branch	1ch-branch	8ch-branch		
1-ch	-	-	ArrayRaw (CH1)	n/a	33.39	36.41
1-ch	-	-	BeamformIt	n/a	31.10	33.21
8-ch	ArrayRaw (CH1)	ArrayRaw	ArrayRaw (CH1)	ArrayRaw	31.81	34.91
8-ch	ArrayRaw (CH1)	ArrayRaw	BeamformIt	ArrayRaw	30.72	33.14
8-ch	ArrayRaw (CH1) + BeamformIt	ArrayRaw	ArrayRaw (CH1)	ArrayRaw	32.02	35.01
8-ch	ArrayRaw (CH1) + BeamformIt	ArrayRaw	BeamformIt	ArrayRaw	**30.12**	**32.33**

Table 5. *Comparison of WERs (%) on AMI Corpus. Note that the conventional multi-channel AMs could not be used with an additional SE module.*

AM	Channels of AM	Speech Enhancement	Dev	Eval
TDNN-BLSTM [16]	Single	-	37.0	40.4
TDNN-BLSTM [16]	Single	BeamformIt	34.2	36.6
Attention-LSTM [15]	Multi	-	35.5	41.0
CNN3D-TDNN-LSTM [28]	Multi	-	32.6	35.4
Ours	Single+Multi	-	32.0	35.0
Ours	Single+Multi	BeamformIt	**30.1**	**32.3**

Table 6. *WERs (%) with different features for 8-ch input branch for AMI (w/o BeamformIt).*

AM	Feature for 8-ch input branch		Dev	Eval
	log-amp.	phase-diff.		
1-ch	n/a	n/a	33.39	36.41
8-ch		$\surd$	32.70	35.78
8-ch	$\surd$		32.13	35.17
8-ch	$\surd$	$\surd$	**32.02**	**35.01**

3.2. Experiments on AMI meeting corpus

3.2.1. Experimental settings

As a second experiment, we evaluated our models by using the AMI Meeting Corpus [42]. The AMI Corpus contains about 100 hours of meeting recordings from four participants. The recordings were conducted using individual headset microphones and an 8-ch microphone array simultaneously. We prepared training, development, and evaluation data by using scripts in the Kaldi toolkit. The AM architecture was the same with the settings used in CHiME-5 except that there were 8 microphones instead of 4.

We first trained a 1-ch AM (AM without N-ch input branch) by using individual headset data ("Headset") as well as the array's first channel data ("ArrayRaw") and array data processed by BeamformIt ("BeamformIt"). We applied speed, volume, reverberation, noise, and bandpass perturbation ($\times 12$) to the "Headset" training data, and we applied speed, volume, and bandpass perturbation ($\times 6$) to the "ArrayRaw" and "BeamFromIt" training data. WERs for this 1-ch AM with and without BeamformIt are shown in the first two rows of Table 4.

Next, we trained the 8-ch AM starting from the 1-ch AM with a randomly initialized multi-channel input branch. In this step, we updated only the parameters of the 8-ch input branch based on the LF-MMI. The results are shown in the lower part of Table 4. Here, we compared the case when we trained the 8-ch AM by feeding only "ArrayRaw" data into the 1-ch input branch and the case when we fed "ArrayRaw" and "BeamformIt" data into the 1-ch input branch. The best results were obtained when we fed the "ArrayRaw" and "BeamformIt" data into the 1-ch input branch in training, and BeamformIt was applied to the 1-ch input branch in decoding. Our model finally achieved a 30.12% WER for the development set and a 32.33% WER for evaluation set, both of which were the best results ever reported for the AMI Corpus to the best of our knowledge (Table 5).

As a supplemental experiment, we evaluated the features for the 8-ch input branch, the results of which are presented in Table 6. As with the CHiME-5 evaluation, we confirmed that both log-amplitude and phase-difference features contributed to the improvements while the log-amplitude feature was the main source of improvement. Finding a better way to utilize phase information is our important future work.

4. CONCLUSION

In this paper, we proposed heterogeneous-input multi-channel acoustic modeling in which AM has both single-channel and multi-channel input branches. In our proposed training pipeline, a single-channel AM was trained first, then a multi-channel AM was trained starting from the single-channel AM with a randomly initialized multi-channel input branch. Our model uniquely uses a complemental SE module while preserving the effectiveness of joint SE and AM training, the effectiveness of which was confirmed with CHiME-5 and AMI experiments.

5. REFERENCES

[1] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, 2011, pp. 437–440.

[2] George E Dahl, Dong Yu, et al., "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[3] Geoffrey Hinton, Li Deng, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[4] Wayne Xiong, Jasha Droppo, et al., "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[5] George Saon, Gakuto Kurata, et al., "English conversational telephone speech recognition by humans and machines," *Proc. INTERSPEECH*, pp. 132–136, 2017.

[6] Dario Amodei, Sundaram Ananthanarayanan, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML*, 2016, pp. 173–182.

[7] Naoyuki Kanda, Yusuke Fujita, and Kenji Nagamatsu, "Lattice-free state-level minimum Bayes risk training of acoustic models," in *Proc. INTERSPEECH*, pp. 2923–2927.

[8] Takuya Yoshioka, Hakan Erdogan, et al., "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. INTERSPEECH*, 2018, pp. 3038–3042.

[9] Jon Barker, Shinji Watanabe, et al., "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.

[10] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc. ASRU*, 2013, pp. 309–314.

[11] Karel Veselỳ, Arnab Ghoshal, et al., "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.

[12] Hang Su, Gang Li, et al., "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," in *Proc. ICASSP*, 2013, pp. 6664–6668.

[13] Daniel Povey, Vijayaditya Peddinti, et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI," *Proc. INTERSPEECH*, pp. 2751–2755, 2016.

[14] Naoyuki Kanda, Yusuke Fujita, and Kenji Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence," in *Proc. ASRU*, 2017, pp. 69–76.

[15] Yu Zhang, Pengyuan Zhang, and Yonghong Yan, "Attention-based LSTM with multi-task learning for distant speech recognition," *Proc. INTERSPEECH*, pp. 3857–3861, 2017.

[16] Vijayaditya Peddinti, Yiming Wang, et al., "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[17] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[18] Hakan Erdogan, John R Hershey, et al., "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016, pp. 1981–1985.

[19] John R Hershey, Zhuo Chen, et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[20] Dong Yu, Morten Kolbæk, et al., "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[21] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Alternative objective functions for deep clustering," in *Proc.ICASSP*, 2018, pp. 686–690.

[22] Katerina Zmolikova, Marc Delcroix, et al., "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. INTERSPEECH*, 2017.

[23] Marc Delcroix, Katerina Zmolikova, et al., "Single channel target speaker extraction and recognition with speaker beam," in *ICASSP*, 2018, pp. 5554–5558.

[24] Naoyuki Kanda, Rintaro Ikeshita, et al., "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *Proc. CHiME-5*, 2018.

[25] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[26] Xiong Xiao, Shinji Watanabe, et al., "Deep beamforming networks for multi-channel speech recognition," in *Proc. ICASSP*, 2016, pp. 5745–5749.

[27] Tara N Sainath, Ron J Weiss, et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. on ASLP*, vol. 25, no. 5, pp. 965–979, 2017.

[28] Sriram Ganapathy and Vijayaditya Peddinti, "3-D CNN models for far-field multi-channel speech recognition," in *Proc. ICASSP*, 2018, pp. 5499–5503.

[29] Katerina Zmolikova, Marc Delcroix, et al., "Optimization of speaker-aware multichannel speech extraction with ASR criterion," in *Proc. ICASSP*, 2018, pp. 6702–6706.

[30] Tsubasa Ochiai, Shinji Watanabe, et al., "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, pp. 2632–2641.

[31] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[32] Yoshua Bengio, Jérôme Louradour, et al., "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.

[33] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117.

[34] Tom Ko, Vijayaditya Peddinti, et al., "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[35] Tom Ko, Vijayaditya Peddinti, et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[36] Daniel Povey, Arnab Ghoshal, et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[37] Shinji Watanabe, Takaaki Hori, et al., "ESPnet: End-to-end speech processing toolkit," *Proc. INTERSPEECH*, pp. 2207–2211, 2018.

[38] George Saon, Hagen Soltau, et al., "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[39] Yiming Wang, Vijayaditya Peddinti, et al., "Backstitch: Counteracting finite-sample bias via negative steps," *Proc. INTERSPEECH*, pp. 1631–1635, 2017.

[40] Tomohiro Nakatani, Takuya Yoshioka, et al., "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. on ASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[41] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[42] Jean Carletta, Simone Ashby, et al., "The AMI meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.