Toshiba's Speech Recognition System for the CHiME 2020 Challenge

Cătălin Zorilă[1], Mohan Li[1], Daichi Hayakawa[2], Min Liu[3], Ning Ding[2] and Rama Doddipatla[1]

[1]Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom
[2]Toshiba Corporation Corporate R&D Center, Kawasaki, Japan [3]Toshiba China R&D Center, Beijing, China
{catalin.zorila, mohan.li, rama.doddipatla}@crl.toshiba.co.uk
{daichi1.hayakawa, ning.ding}@toshiba.co.jp, liumin@toshiba.com.cn

Abstract

This paper summarizes the Toshiba entry for Track 1 of CHiME 2020 challenge, corresponding to the multi-array speech recognition task. The system is based on conventional acoustic modeling (AM), where phonetic targets are tied to features at the frame-level, and it consists of a combination of convolutional neural networks (CNNs) (with or without residual connections) and factorized time delay neural networks (TDNNFs). We also explored several enhancement strategies for the train and test data, speaker normalization and discriminative training. Results are reported using the provided 3-gram language model (3G LM) and after rescoring with a neural network language model (RNN LM). Following system combination, the submitted system achieves a performance of 35.89% and 37.54% word error rate (WER) using 3G LM on the development (DEV) and evaluation (EVAL) sets, respectively. Using the RNN LM, our system achieves a performance of 34.83% and 36.83% WER on DEV and EVAL, respectively. Proposed system was ranked 4th in both the constrained and the unconstrained language model subtracks.

1. Introduction

The Toshiba entry for Track 1 of CHiME 2020 challenge [1] is presented here. The goal is on building an automatic speech recognition (ASR) system where the speaker diarization information is provided. Track 1 is a follow-up of the multi-array track from CHiME 2018 challenge [2], and is ranked into two categories (A or B), based on the type of acoustic modeling architecture and the type of language model used. The system presented here uses conventional acoustic models trained with phonetic targets tied to features at frame-level. Results are presented using both the baseline 3-gram language model as well as rescoring with a neural network LM. The sections below describe the system's components.

The CHiME 2020 corpus is the same as the CHiME 2018 (CHiME-5) one, and consists of twenty dinner party recordings (sessions) lasting about two hours each. There are 16 training sessions, 2 sessions for development and 2 sessions for evaluation. A session has four participants and is divided into three parts (kitchen, dining room and living room), corresponding to the location from where the audio data is collected. Conversations in each location are captured by: (a) two distant 4-channel input microphone arrays, and (b) four in-ear (worn, one per participant) binaural microphones. The worn microphone recordings were used to annotate the data. There are six distant microphone arrays (24-channels) and four in-ear stereo microphones (8-channels) in total. None of the devices were synchronized during recording, therefore a correlation based approach was initially employed to align data [2]. In addition to the synchronization problem, speech is also corrupted by noise, reverberation and overlapping speakers [3].

A more accurate array synchronization method was provided for CHiME 2020 [1]. Also, the baseline acoustic model and enhancement pipeline were updated to match the state-of-the-art results reported in [4–6]. The factorized time delay neural networks [7], the multi-channel Guided Source Separation (GSS) enhancement [8], and cleaning up both the training and test datasets [6] were proven to be quite effective in reducing the word error rate.

The rest of the paper is structured as follows. Section 2 summarizes the speech enhancement strategies used for our system, Section 3 describes the proposed acoustic model topology and training strategy, Section 4 presents the proposed neural network based language model, and Section 5 shows the recognition accuracies of our final ASR system. Finally, the paper is concluded in Section 6.

2. Speech enhancement

Dereverberation and source separation were applied to process the speech signals for our system. Both are briefly summarized below.

2.1. Dereverberation

Reverberation pose a significant challenge for distant-talking speech recognition and requires dedicated solutions to alleviate its effects [9]. It has been shown that reducing the late reverberation (impulse response shortening) prior to feature extraction improves considerably the recognition accuracy. The multiple input multiple output (N inputs and N outputs) version of the Weighted Prediction Error (WPE) method was used to achieve room impulse shortening for our submission [10, 11][1].

2.2. Guided Source Separation

GSS is a blind source separation method aimed to reduce the effect of speaker overlap that was initially proposed for CHiME-5 [8]. Due to its success in reducing the word error rate for CHiME-5 [5, 6], the vanilla (one-stage) multi-array GSS implementation was released with the baseline system for CHiME-6 Track 1. Block diagram of one-stage GSS is depicted in Fig. 1 (switch K=1), and is concisely described next.

First, the time-frequency (T-F) representation of N-channel mixture of reverberated overlapped speech is obtained using the Short Term Fourier Transform (STFT), and the late reverberation is removed by means of WPE. Then, the parameters of a spatial mixture model are estimated using the Expectation Maximization (EM) algorithm, and posterior probabilities of a speaker being active are computed. The posteriors (speaker masks) are employed to estimate spatial covariance matrices of target and interference speakers, and then derive the steer-

[1]http://github.com/fgnt/nara_wpe

ing vector of a Minimum Variance Distortionless Response (MVDR) beamformer [12]. Audio waveform is reconstructed using the inverse STFT and the overlap-and-add technique.

For the one-stage GSS, the EM algorithm is initialized using the speaker diarization information inferred from the original CHiME-6 transcription. Previous experiments have shown, however, that refining the EM initialization using voice activity detection information provided by an ASR system can significantly improve the recognition accuracy, as described in [6]. Therefore, the test data are firstly enhanced using GSS (K=1 in Fig. 1), then they are transcribed using a pre-trained ASR system, and, finally, the silence information are used to improve the EM initialization for a second run of GSS (K=2 in Fig. 1). This implementation flavour is referred to as *two-stage GSS*. We have used both one- and two-stage GSS flavours in our final system.

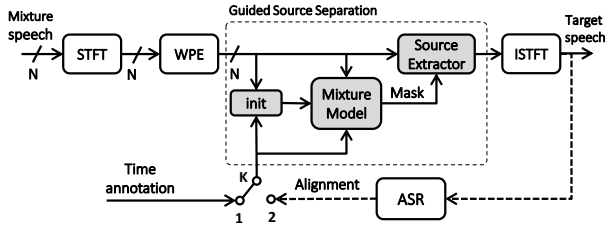


Figure 1: *Block diagram of one- and two-stage Guided Source Separation enhancement.*

3. Acoustic models

3.1. Baseline acoustic model

The baseline acoustic model for CHiME-6 consists of 14 factorized time delay neural network [7] layers, and is trained using unprocessed worn (W) and array (U) speech segments. Simulated reverberated array speech (U.rvb) and 3-fold speed perturbation (SP) are used to augment the training data [4, 13]. The acoustic features are 40-dim MFCC and 100-dim i-vectors [14], and the training is conducted in KALDI using the LF-MMI criterion [15]. During testing, speech is enhanced using one-stage GSS and the i-vectors are refined using a 2-pass decoding approach. For more details, the reader is referred to [1] [2].

3.2. Proposed acoustic models

Several acoustic models were combined to form our final system. They contrasted in both network topology and training strategy.

Concerning the network architecture, the topologies we explored consisted of CNN (with and without residual connections) and TDNNF blocks, as depicted in Figs. 2 and 3. In Fig. 2, BN and FC denote batch normalization and fully-connected layers, respectively, k is the kernel size, s is the stride in time and frequency, d, b and o are dimensions of FC layers. The linear layers factorized under a semi-orthogonal constraint are marked with an '*' (e.g., Fig. 2c), and the corresponding dimension (b) is referred to as the bottleneck dimension [7].

Three distinct topologies are proposed in Fig. 3. One consists of 9 CNN (without residual connections) and 10 TDNNF layers, another one is based on 40 ResNet CNN layers, and the third one is formed of 40 ResNet CNN layers followed by 9 TDNNF layers. An additional CNN-TDNNF topology (not depicted in Fig. 3) was also employed, whose structure is similar with that in Fig. 3a, but had the first CNN layer removed. The dropout rate was set to zero for all experiments, and n refers to the context size.

[2] http://github.com/kaldi-asr/kaldi/tree/master/egs/chime6/s5_track1

Table 1: *Configuration of AMs used in this paper.*

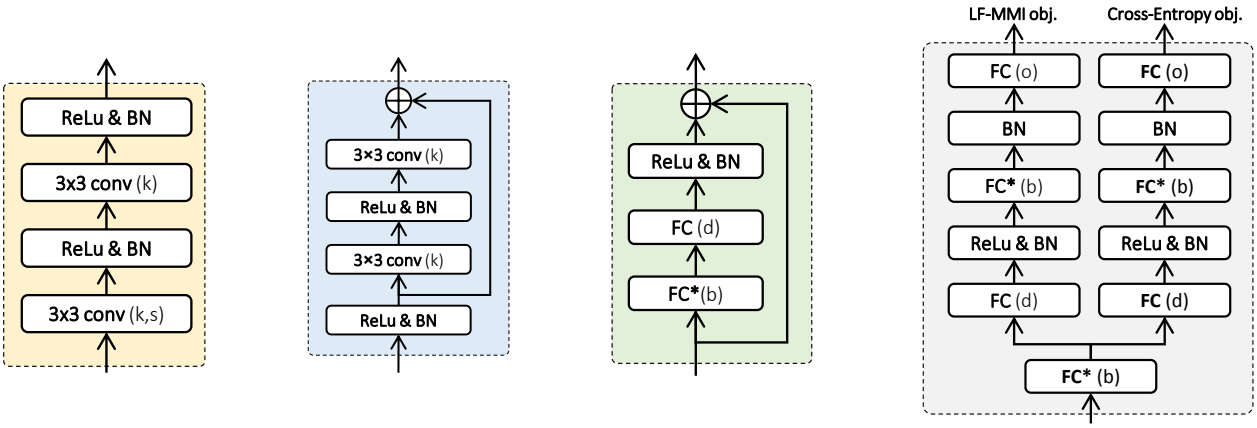
Enh. in train	Topology	SID	DT	VTLN	2-pass	hrs
W+U+U.rvb	TDNNF(15)	Base			yes	1407
W+U.WPE	CNN-TDNNF(19)	A			no	802
W+U.GSS12	CNN-TDNNF(18)	B			no	102
	CNN-TDNNF(19)	C	✓		no	309
		D	✓	✓	no	309
		E			yes	309
		F	✓		yes	309
		G		✓	yes	309
		H	✓	✓	yes	309
	RESNET(40)	I			yes	309
		J	✓		yes	309
		K		✓	no	309
	RESNET-TDNNF(49)	L			no	309
		M	✓		no	309

Acronyms:

2-pass	2-pass decoding for i-vector refinement
CNN	Convolutional Neural Network (without residual connections)
DT	Discriminative Training
RESNET	CNN with residual connections
SID	System ID
TDNNF	Factorized Time Delay Neural Network
U	array data
U.GSS12	Guided Source Separation (12-ch) enhanced U data
U.rvb	simulated reverberated U data
U.WPE	dereverberated U data using WPE [11]
VTLN	Vocal Tract Length Normalization
W	worn uprocessed data

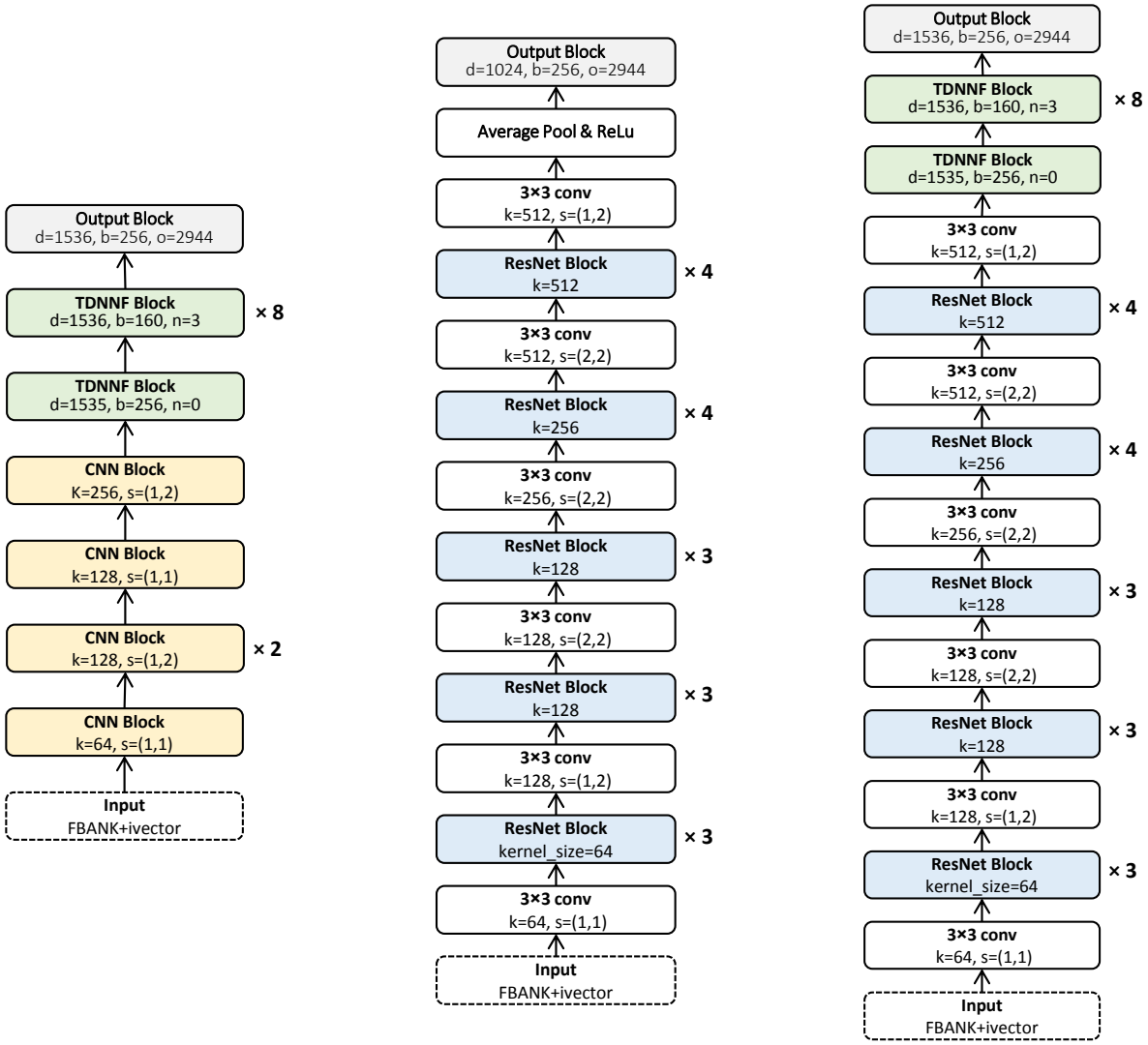
Concerning the training strategy, all models were trained with unprocessed worn and enhanced array data, and the signal processing was either WPE dereverberation [11] or one-stage multi-channel GSS (12-channel input, GSS12). Two-stage GSS with 12 and/or 24-channel input was used during test, as in [6]. Models were initially trained using LF-MMI criterion, and were later refined using discriminative training (DT) [16]. Vocal tract length normalization (VTLN) [17] was also applied, with warp factors estimated for each speaker, room and session. I-vector models were trained on both unwarped and warped acoustic features. Furthermore, the test i-vectors were refined using 2-pass decoding.

Table 1 shows the configuration of all models used for the final system. As shown later, the best performance is achieved with discriminative training, VTLN normalization and 2-pass decoding. 64-dim FBANK and 100-dim i-vectors were the acoustic features used for all models but system ID (SID) B, where 64-dim FBANK were combined with 10-dim excitation based features in [18]. The excitation based features are designed to partially recover the glottal source information lost during FBANK computation. In term of data augmentation, except SID B, all models in Table 1 used 3-fold speed perturbation. Training and decoding were performed in KALDI. Last column of Table 1 presents the total amount of training data (in hours) for each acoustic model. Notably, most acoustic models were trained with roughly 300 hrs of data, which is a fraction of that used to train the baseline model (about 1400 hrs).



(a) *CNN Block* (b) *ResNet CNN Block* (c) *TDNNF Block* (d) *Output Block*

Figure 2: *Structure of basic network blocks of proposed acoustic models. See Section 3.2 for details.*



(a) *CNN-TDNNF(19 layers)* (b) *RESNET(40 layers)* (c) *RESNET-TDNNF(49 layers)*

Figure 3: *Proposed acoustic model topologies.*

4. Neural network language model

The proposed model consisted of two Long-Short Term Memory (LSTM) recurrent neural network [19] and three TDNN layers connected as in Fig. 4. LSTM layers had a cell dimension of 800, with a recurrent projection of 256 and a non-recurrent projection of 128; the word embedding dimension was 800. The left and right context size (n) of TDNNs is specified in Fig. 4. Parameters were chosen to reduce the perplexity on the development set of CHiME-6 (Table 2).

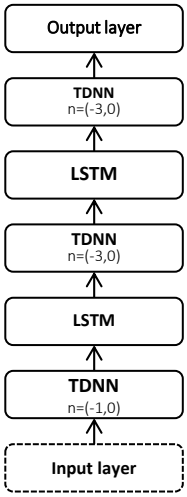


Figure 4: *Proposed TDNN-LSTM language model.*

Table 2: *Perplexity of baseline and proposed language models.*

Language model	Perplexity
Baseline (3-gram)	154.2
Proposed (TDNN-LSTM)	140.5

5. System combination

The individual and combined performances of systems in Table 1 were assessed for the constrained (Category A) and unconstrained (Category B) language models. Results are presented below.

5.1. Results Category A

Performance of individual systems described in Table 1 using the baseline 3-gram language model is presented in Table 3. Performance of baseline CHiME 2020 Track 1 system is also included for comparison purposes; $^{*}*$note that Base numbers reported in Table 3 are with one-stage GSS12 enhanced test data (without ASR refinement).

From the table, one can observe a significant ASR accuracy improvement relative to the Base model for our systems. This result is remarkable since our models were trained using a fraction of the data necessary to prepare the baseline model (see last column in Table 1). The best accuracy is achieved by the RESNET topology, and performing discriminative training (SID J) and vocal tract length normalization (SID K) helps reduce the WER further. Our experiments have shown that performing lattice combination using GSS12+ASR and GSS24+ASR streams lead to significant WER improvements, therefore both streams have been included in the submission system. Lattice combination of all systems and streams in Table 3 yielded 35.89% and 37.54% WER on DEV and EVAL, respectively. This accuracy was ranked 4th in the official Category A results from the challenge.

Table 3: *Performance in %WER of individual components of final system for 3G LM.*

SID	DEV		EVAL	
	GSS12+ASR	GSS24+ASR	GSS12+ASR	GSS24+ASR
Base	51.39$^{*}*$	-	51.38$^{*}*$	-
A	45.81	44.79	46.09	46.78
B	44.88	-	49.48	-
C	42.47	42.15	44.28	45.03
D	41.74	41.73	43.84	44.92
E	42.67	42.28	44.78	45.11
F	41.78	41.49	44.21	44.72
G	41.66	41.13	44.11	44.66
H	41.27	41.34	43.71	44.80
I	41.34	41.06	42.42	43.09
J	41.07	40.55	41.84	42.94
K	40.86	40.53	42.41	43.12
L	42.03	-	42.78	-
M	41.71	-	42.86	-
A-M	**35.89**		**37.54**	

Table 4: *Performance in %WER of individual components of final system for RNN LM.*

SID	DEV		EVAL	
	GSS12+ASR	GSS24+ASR	GSS12+ASR	GSS24+ASR
A	44.64	43.42	44.62	45.38
B	43.39	-	45.66	-
C	41.01	40.71	43.07	43.92
D	40.53	40.56	42.96	43.74
E	41.42	40.93	42.85	43.74
F	40.74	40.38	42.84	43.73
G	40.57	39.99	42.74	43.43
H	40.10	40.02	42.70	43.99
I	40.29	40.04	41.84	42.29
J	40.11	39.76	41.00	42.19
K	39.94	39.62	41.42	42.27
L	41.14	-	42.05	-
M	40.85	-	42.16	-
A-M	**34.83**		**36.83**	

5.2. Results Category B

Similar results for the RNN language model are depicted in Table 4. Best recognition accuracy was achieved by the RESNET architecture with vocal tract length normalization (SID K). Combining lattices of all systems and streams yielded 34.83% and 36.83% WER on DEV and EVAL, respectively. This accuracy was ranked 4th in the official Category B results from the challenge.

5.3. Discussion

A session and room breakdown performance of the final systems for the 3-gram and RNN language models is provided in Table 5. Notably, except for session S09, the KITCHEN condi-

tion of all other sessions had much worse performance than the DINING and LIVING conditions. The performance gap can be attributed to the specific noise characteristics and levels found in the kitchen environment, or may be due to the higher rate of speaker movement during the food preparation phase of the dinner party. This issue will be addressed in our future work.

Table 5: *Detailed % WER performance of the final system.*

Session	Room	3G LM		RNN LM	
		DEV	EVAL	DEV	EVAL
S02	DINING	39.84	-	38.46	-
	KITCHEN	41.67	-	40.65	-
	LIVING	32.65	-	31.83	-
S09	DINING	36.03	-	34.47	-
	KITCHEN	33.63	-	32.62	-
	LIVING	31.14	-	30.14	-
S01	DINING	-	31.31	-	30.56
	KITCHEN	-	53.03	-	52.80
	LIVING	-	43.38	-	42.80
S21	DINING	-	29.91	-	28.64
	KITCHEN	-	45.45	-	44.99
	LIVING	-	30.16	-	29.25
Overall		**35.89**	**37.54**	**34.83**	**36.83**

6. Conclusion

In this paper we have summarized the Toshiba entry for Track 1 of CHiME 2020 Challenge. A conventional HMM-DNN ASR system was proposed, consisting of a combination of CNN and TDNNF acoustic model topologies, two-stage multi-array GSS enhancement, speaker normalization using VTLN and second pass discriminative training. Results were reported using the 3-gram LM provided by the organizers and after rescoring with an RNN LM. For the 3-gram LM, our system has achieved 35.89% and 37.54% WER on the development and evaluation sets, respectively. For the RNN LM, our system has achieved 34.83% and 36.83% WER on the development and evaluation sets, respectively. The system was ranked 4th in both the constrained and the unconstrained language model subtracks.

7. References

[1] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth CHiME speech separation and recognition challenge: dataset, task and baselines," in *Proc. Interspeech*, Sep 2018, pp. 1561–1565.

[3] C. Zorilă and R. Doddipatla, "On reducing the effect of speaker overlap for CHiME-5," in *Proc. ICASSP*, April 2019, pp. 6645–6649.

[4] V. Manohar, S.-J. Chen, Z. Wang, Y. Fujita, S. Watanabe, and S. Khudanpur, "Acoustic modeling for overlapping speech recognition: JHU Chime-5 Challenge system," in *Proc. ICASSP*, May 2019, pp. 6665–6669.

[5] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, and R. Haeb-Umbach, "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR," in *Proc. Interspeech*, Sep 2019.

[6] C. Zorilă, C. Boeddeker, R. Doddipatla, and R. Haeb-Umbach, "An investigation into the effectiveness of enhancement in ASR training and test for CHiME-5 dinner party transcription," in *Proc. ASRU*, 2019, pp. 47–53.

[7] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018, pp. 3743–3747.

[8] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. of CHiME-5 Workshop*, 2018.

[9] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[10] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[11] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Proc. of ITG Fachtagung Sprachkommunikation*, Oct 2018.

[12] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[15] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[16] D. Povey, "Discriminative training for large vocabulary speech recognition," *Ph.D. thesis*, 2004.

[17] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," *Technical Report CMU-CS-97-148*, 1997.

[18] T. Drugman, Y. Stylianou, L. Chen, X. Chen, and M. Gales, "Robust excitation-based features for automatic speech recognition," in *Proc. ICASSP*, 2015, pp. 4664–4668.

[19] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.