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@crl.toshiba.co.uk

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% 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.

1. System Description

The Toshiba system entry for Track 1 of CHiME 2020 challenge is presented here. Track 1 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 [1], 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.

1.1. Speech Enhancement: Guided Source Separation

GSS enhancement is a blind source separation method aiming to reduce the effect of speaker overlap initially proposed in [2]. Given a mixture of reverberated overlapped speech, GSS estimates the parameters of a spatial mixture model which is then used to calculate the posterior probabilities of active speakers to drive a mask-based beamforming. Both vanilla (baseline) multi-array GSS implementation provided by the organizers and a modified version have been used to enhance speech for our system. The modified GSS enhancement (GSS+ASR) was based on improving the original speaker diarization information provided by the challenge using ASR-estimated silence information, as described in [3].

1.2. Acoustic Model

A 15-layer factorized time delay neural network topology is proposed as the baseline acoustic model for the challenge. Training data consists of unprocessed worn (W) and array (U) data, augmented with simulated reverberated array speech (U.rvb) and 3-fold speed perturbation (SP) [5]. 40-dim MFCC and 100-dim i-vectors are used as acoustic features, training criterion is LF-MMI and the standard language model is 3-gram. The best accuracy for the baseline model is achieved using GSS enhancement and 2-pass decoding with i-vector refinement during testing.

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

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

Acronyms:
2-pass 2-pass decoding with 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 [4]
VTLN Vocal Tract Length Normalization
W worn uprocessed data

For our system we have explored variability in: (a) acoustic model topology; (b) training and test data enhancement, and (c) acoustic features and speaker adaptation. Several AM topologies were chosen consisting of CNNs (with and without residual connections) and TDNNFs (Table 1). Training data and number of AM layers are also specified in Table 1, as well as whether discriminative training (DT) is applied on top of standard LF-MMI trained models, speaker normalization is active (vocal tract length normalization, VTLN), or whether 2-pass decoding with i-vector refinement is being computed. All models were trained using unprocessed worn data and array speech enhanced either using Weighted Prediction Error (WPE) dereverberation [6, 4] or vanilla multi-channel GSS enhancement (12-channels, GSS12). The test data were enhanced using GSS+ASR with 12 and/or 24-channels [3]. Acoustic features were 64-dim filter-bank (FBANK) features and 100-dim i-vectors for all models but system ID (SID) B, where 64-dim

Table 2: *Performance in %WER of individual components of final system for the baseline language model.*

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**	

FBANK were combined with 10-dim excitation based features in [7]. Except SID B, all systems have used 3-fold SP; the training and decoding were performed in KALDI.

1.3. Neural network language model

A TDNN-LSTM language model has been used for performing language model rescoring. The network consists of two LSTM layers interleaved between 3 TDNN layers. The LSTM layer has a cell dimension of 800, with a recurrent projection of 256 and a non-recurrent projection of 128; the word embedding dimension is 800. This model yields a perplexity of 140.5.

2. Results Category A

Performance of individual systems described in Table 1 using the baseline 3-gram language model is presented in Table 2. Performance of baseline CHiME 2020 system is also included for comparison purposes; *note that Base numbers reported in Table 2 are with GSS12 processed test data (without ASR refinement).

From the table, one can observe a significant ASR accuracy improvement relative to the Base model for our systems. 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 have been included in the final system. Lattice combination of all systems and streams in Table 2 yielded 35.89% and 37.54% WER on DEV and EVAL, respectively.

3. Results Category B

Similar results for the RNN language model are depicted in Table 3. Combining lattices of all systems and streams yielded 34.83% and 36.83% WER on DEV and EVAL, respectively.

4. Performance analysis

A session and room breakdown performance of the final systems for the 3-gram and RNN language models is provided in Table 4.

Table 3: *Performance in %WER of individual components of final system for the RNN language model.*

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**	

Table 4: *Detailed % WER performance for 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**

5. 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 CNNs and TDNNFs acoustic model topologies, ASR-refined multi-array GSS enhancement, speaker normalization using VTLN and second pass discriminative training. Results were reported using provided 3-gram LM 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.

6. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.