The Toshiba Entry to the CHiME 2018 Challenge

Rama Doddipatla[*]*, Takehiko Kagoshima*[†]*, Cong-Thanh Do*[*]*, Petko N. Petkov*[*]*, Cătălin Zorilă*[*]*,*
Euihyun Kim [†]*, Daichi Hayakawa*[†]*, Hiroshi Fujimura*[†] *and Yannis Stylianou*[*]

[*]Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom
[†]Toshiba Corporation Corporate R&D Center, Kawasaki, Japan

[*]firstname.lastname@crl.toshiba.co.uk, [†]firstname.lastname@toshiba.co.jp

Abstract

This paper summarises the Toshiba entry to the single-array track of the CHiME 2018 speech recognition challenge. The system is based on conventional acoustic modelling (AM), where phonetic targets are tied to features at the frame-level, and use the provided tri-gram language model. The system is ranked in category A that focuses on acoustic robustness. Array signals are first enhanced using speaker dependent generalised eigenvalue (GEV) based beamforming. Two different acoustic representations are then extracted from the enhanced signals: i) log Mel filter-bank and ii) subband temporal envelope (STE) features. Separate acoustic models, trained on each set, are used for lattice combination. The AM combines convolutional and recurrent architectures in a single CNN-BLSTM model. Speaker adaptation, limited to vocal tract length normalisation (VTLN), de-reverberation and speaker suppression are also considered. Following system combination, the Toshiba entry achieves 60.8% word error rate (WER) on the development (*dev*) set and 56.5% WER on the evaluation (*eval*) set respectively. The system is ranked 4^{th} in the A category.

1. Introduction

CHiME 2018 targets distant conversational ASR using microphone arrays in everyday home environments. The challenge had a single array track and a multiple array track. The single array track specifies the reference array to use the test data from. On the other hand, data from all the arrays can be used at test time in the multiple array track. The submitted systems are ranked in categories A or B based on the type of acoustic and language models used for building the ASR system. If conventional frame-based acoustic models are used along with the provided tri-gram language model, the systems are ranked in category A. The systems in category B have no restrictions: they can explore any acoustic and language model and can also use additional external training data. More details about the training data and the challenge instructions can be found in [1]. The Toshiba entry to the challenge focuses on reference array track and uses conventional frame-based AM and the provided LM, hence will be classed into category A. The rest of the paper is organised as follows: first, the paper describes the system's components, that include enhancement, front-end, speaker adaptation and system combination, and then presents how these components were combined to reach the final submission system.

2. Baselines from the Challenge

The Challenge organisers provided a baseline system developed in KALDI [2]. A time-delay neural network (TDNN) [3] is used as the AM, which is trained using lattice-free (LF-) MMI [4]. The TDNN AM has 8 ReLU batch normalisation layers with the following context on each of the layers: [-2:2],0,[-1:1],0,[-1:1],0,[-3,0,3],[-3,0,3],[-6,-3,0]. The first entry shows the context used on the input layer (2 frames to the left and 2 frames to the right including the centre frame). The models are trained using the chain framework in KALDI. A lexicon and language model are also provided by the challenge organisers [1]. The baseline AM is trained using data from the worn (W) microphones and randomly chosen 100k utterances from the all the available arrays (U). The system performs speed-perturbation (sp) [5] to increase the training data by three folds and includes *i-vectors* [6] to perform speaker adaptation. The performance of the baseline system is presented in Table 1.

Table 1: *Baseline performance (%WER) of TDNN acoustic models on CHiME5 on the* dev *set.*

CHiME5	GMM-HMM	TDNN
worn (W)	71.8	47.7
array (U)	91.2	80.8

The GMM-HMM is a speaker adaptive training (SAT) model using feature space maximum likelihood linear regression (FMLLR). Though the Challenge only ranks the performance of the system based on the %WER's on the array data, the performance of the worn data is also presented to show the complexity of the task. One can observe that the performance of W data is already close to 50% WER. The performance of U data gets worse compared to the W data as we move from a close talk to a distant speech recognition task.

2.1. Effect of Speed perturbation and i-vectors

An investigation is performed to understand the contribution of speed-perturbation and the use of *i-vectors* towards ASR performance. The performance using the TDNN AM is presented in Table 2.

Table 2: *Performance (%WER) on TDNN AM using speed-perturbation and* i-vectors *on the array data.*

dev-U	sp	i-vectors	% WER
	-	-	83.8
	-	+	80.9
	+	+	80.8

One can observe that turning off *sp* did not have much effect on the ASR performance of the U data, while turning off the *i-vectors* seems to degrade the performance. So for all initial investigations, *sp* is turned off (for a faster turn around time during AM training), but *i-vectors* are used into the system development pipeline.

3. Acoustic models

A variety of AMs were investigated to see if they have an advantage over the baseline TDNN AMs. In this direction, we explored both uni-directional and bi-directional long short-term memory networks (LSTM) [7]. Convolutional neural networks (CNN) [8] in combination with LSTMs were also explored. The performance results of different architectures are presented in Table 3. For all the experiments presented here, *sp* is turned off and *i-vectors* are used for speaker adaptation.

Table 3: *Performance (%WER) comparison of different AM architectures.*

W+U100k	*dev*-U
TDNN	80.9
LSTM	76.9
BLSTM	76.6
CNN-LSTM	75.3
CNN-BLSTM	74.9

From Table 3, one can observe the progression of the change in performance for U data. All the acoustic models are trained using W+100k data. The CNN-BLSTM AM consisting of 2 CNN layers at the front followed by 3 BLSTM layers provides the best performance. The CNN layers are 2D-CNN layers with 3x3 filter kernels and having 256 and 128 filters respectively. Each BLSTM layer has a cell dimension of 1024 and recurrent projection of 256. The *i-vectors* on the input are bypassed from the CNN layers and presented along with the output of CNN to the BLSTM layers. A context of 40 frames (including both left and right) is used for the BLSTM layers. Based on these results, the CNN-BLSTM AM was chosen to be the AM for our submission system.

4. Front-end: FBANK and STE

Two types of acoustic features were used: log-Mel filter-bank (FBANK) and subband temporal envelope (STE) [9] features. In both cases 40 coefficients are extracted per frame. In contrast to conventional FBANK features which extract information from spectral domain, STE features extract information from slowly-varying temporal envelopes in the frequency subbands of speech. In this respect, speech signal is filtered with a Gammatone filter-bank. STEs are then extracted by applying full-wave rectification and low-pass filtering, with a cut-off frequency of 50 Hz, to the subband signals. The features coefficients are then computed from overlapping frames of the STEs [9]. The STE features extraction pipeline is depicted in Fig. 1. More details about these features can be found in [9]. Both FBANKs and STEs are mean-normalised on a per-segment basis. The motivation for using two feature sets is the expected gain from combining complementary information. The performance of both the features are presented in Table 4. One

Table 4: *Performance (%WER) of FBANK and STE features.*

CNN-LSTM	FBANK	STE
W+U100k	75.3	75.1

can observe that both features have similar accuracies and performing lattice combination can be advantageous. Though most of the investigations are presented using FBANK features, the final submission systems will have a combination of systems

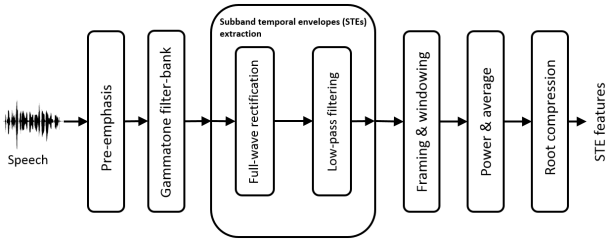

Figure 1: *Subband temporal envelope (STE) features extraction pipeline [9].*

trained using both FBANK and STE features. Unless specified, all the results presented in the discussion will be either MFCC or FBANK features.

5. Separate AMs for each array

Instead of using data from all the arrays, separate AMs are trained using data from each array. The motivation is to avoid the problem of synchronisation across arrays and perform system combination on the ASR outputs of each of these systems. The AMs are trained using data from a specific array along with the W data. A CNN-BLSTM AM architecture is employed. The performance on the *dev* set is presented in Table 5.

Table 5: *Performance (%WER) of AMs trained using data from specific arrays.*

CNN-BLSTM	no-sp	sp
W+U100k	74.9	-
W+U01	-	71.4
W+U02	73.8	71.7
W+U04	73.8	70.7
W+U05	74.2	72.6
W+U06	73.7	72.1
W+Uall	-	70.1

One can observe that the performance of individual arrays is better than training the AM with a mix of data from all the arrays. The amount of training data from each array is more than W+U100k and could be a possible reason for better performance. It is interesting to note that the performance on the *dev* set is very similar for all the AMs trained using data from specific arrays. Adding *sp* data into training the AM further improved the performance. For the rest of the experiments presented in the paper, *sp* is always used for AM training. The AM trained using W+Uall (worn and all the data available from the arrays) seem to perform the best, also indicating that having access to more training data improves the AM.

6. Speech enhancement

6.1. Speaker dependent GEV beamforming

A variant of neural net (NN) supported GEV-beamforming in STFT domain is first applied [10, 11]. The objective is to enhance the target speaker while suppressing background interference (noise and competing speakers). The time-frequency masks for speech and noise are estimated from the input speech using an NN model. These are subsequently modified using speaker identity information (Fig. 2). The dominant speaker in each frame was estimated using a GMM classifier (speaker clus-

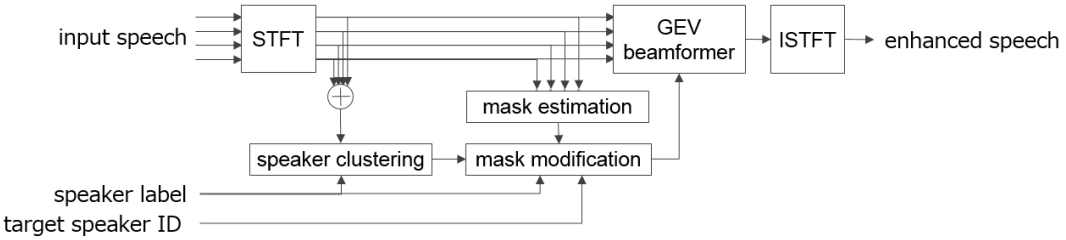

Figure 2: *Proposed speaker dependent GEV-based speech enhancement.*

tering), while the speaker labels were extracted from the transcriptions. The speech mask was set to zero in frames where the dominant and the target speaker differ, and the noise mask was set to one where competing and target speakers overlap.

The NN model was trained with worn microphone data from the train set of CHiME-5. Single speaker segments were collected, and noise reduction using L and R channels was applied to produce the "clean" speech. L channel data of non-speech portions represents noise. Noisy speech data was simulated using the clean speech and the noise, and ideal binary masks for speech and noise were generated for training the NN. The input feature was STFT power spectra with 2827 dimensions (257 frequency bins $\times$ 11 frames). The NN was fully-connected comprising 4 hidden layers (1600 nodes) with sigmoid activations in the output layer and eLU elsewhere. The output layer had 514 nodes (257 frequency bin $\times$ 2 masks). GMM training and clustering were performed using speech data averaged over the microphones of the reference array. The features were 25-dimensional MFCCs, and the GMM had 16 components. Four clusters were used (one for each speaker).

The initial GMM for each speaker was trained with single speaker portions of the signal according to the transcriptions. Thereafter, clustering and updating the models were repeated until convergence. The frames in the single speaker segments were fixed to the cluster of the active speaker. Those in the other segments were clustered to one of the speakers whose utterance labels included the time frame according to the GMM likelihood. Frames in which the normalised likelihood fell below 0.9 were removed from the clusters. A GMM was trained for each reference array using utterances from the given array only.

Table 6: *Performance (%WER) of AMs trained using GEV enhanced data for specific arrays.*

CNN-BLSTM	nosp	sp
W+U01	69.2	67.4
W+U02	67.3	67.0
W+U04	68.1	66.1
W+U05	70.1	66.9
W+U06	68.5	66.3
W+Uall	66.7	64.9

The proposed enhancement is applied both in training and recognition. The ASR performance using the proposed speaker dependent GEV are presented in Table 6. One can observe that speaker dependent GEV helps improve the ASR performance when compared with the case without GEV enhancement (Table 5). The AM trained using data from all the arrays performs best as also observed in Table 5.

6.2. Enhancement using WPE

Physical separation on the order of a few meters between the speakers and the microphone arrays results in noticeable reverberation. Signal de-reverberation prior to feature extraction is, thus, expected to improve recognition performance [12]. A fundamental challenge in this context is that in the presence of overlapped speech the expected gain from de-reverberation is limited. Due to time constraints, reverberation and speaker overlap are treated as separate problems in this system.

The weighted prediction error (WPE) is a prominent approach to reducing late reverberation [13]. It has been used successfully in past speech recognition challenges [14]. The method in its original formulation relies on iterative estimation (based solely on the target utterance) to compute the optimal de-reverberation filter coefficients. It is observed that for short utterances the iterative approach becomes unstable and ineffective. This is problematic in the context of CHiME5, where due to the spontaneous interaction among the participants, a large portion of the utterances are short.

A possible work-around is seen in the use of a neural-network (NN) supported WPE, which improves both stability and performance [15]. A disadvantage in the formulation of this method is the dependence on parallel training data, i.e., reverberant and "clean" (direct sound and early reflections) speech. The close-talk CHiME5 data is not well-suited for use as the target "clean" speech suggesting that an alternative approach to training the supporting NN is needed.

An in-house method for training the supporting NN was developed. The approach is unsupervised as it does not require parallel data or joint training with an acoustic model. Single-channel de-reverberation on top of GEV-enhanced signals was only considered in this context.

Table 7: *Performance (%WER) of AMs using WPE.*

CNN-BLSTM	no WPE	WPE
W+Uall (+sp)	64.9	63.3

The performance of the system is presented in Table 7. One can observe that enhancing the train set with WPE and retraining the acoustic model improves the system performance. Of all the above systems, this system has the best performance.

6.3. Enhancement using Speaker Suppression

The test utterances that were enhanced using GEV could not completely separate overlapping speakers. A mask based speaker suppression approach has been attempted to see if it improves the system performance. The masks were estimated over time using a frame-wise RNN speaker classifier trained on non-overlapping portions of speech extracted from the transcription

files. MFCCs (dimension 24) with delta and delta-delta were used as acoustic features; they were extracted from 32-ms of speech every 16-ms. Speaker predictions were made using the maximum likelihood criterion, followed by 13 taps median filtering. The frames where the target speaker was dominant had gain one, and the frames where the interfering speakers were dominant had gain 0.001. Finally, all mask weights were post-filtered using double exponential smoothing with the attack and release constants of 0.1 and 0.99, respectively. The performance of the proposed approach is presented in Table 8.

Table 8: *Performance (%WER) of AMs using speaker suppression.*

CNN-BLSTM	no SS	SS
W+Uall (+sp)	64.9	64.8

7. Speaker adaptation using VTLN

VTLN is a simple technique for speaker adaptation [16]. It scales the frequency axis linearly to normalise variability in the speech signals caused by speaker differences. The scaling factor is estimated using a grid search in the range from 0.85 to 1.25 in steps of 0.01. The warped features, derived based on the optimal scale factors, are used both during training and recognition. Estimation of the VTLN warp factors for the *dev* and *eval* sets requires a two-pass approach. The performance of VTLN is presented in Table 9.

Table 9: *Performance (%WER) of AMs using VTLN adaptation.*

CNN-BLSTM	no VTLN	VTLN
W+Uall (+sp)	64.9	64.1

From the table, one can observe that VTLN improves the performance of the system already enhanced with GEV. VTLN estimation can be influenced by the overlap present in the input speech. A better speaker separation or suppression approach before VTLN estimation can further help improve the system performance.

8. System Combination

Finally, the system combination of the above components is explored. A summary of the individual performance of various components previously described is presented in Table 10.

Table 10: *Summary of WER (%) on the development set.*

Track	Data	System	FBANK	STE
Single	W + U01		67.4	66.6
	W + U02		67.0	65.8
	W + U04		66.1	66.0
	W + U05		66.9	65.6
	W + U06		66.3	66.7
	W + Uall	C	64.9	-
	W + Uall - SS	D	64.8	-
	W + Uall - VTLN	E	64.1	-
	W + Uall - WPE	F	63.3	-

Systems were are also trained using sub-band temporal envelope (STE) features, other the FBANK features, only using data from the individual arrays. The primary motivation is to perform system combination on the ASR outputs. The Uall systems were developed only for the FBANK features due to time constraints.

The system combination results are presented in Table 11. Lattice combination on the individual arrays either FBANK (A) or STE (B) performed better than the systems trained using data from all the arrays (C). Further combining A and B provided further gains in ASR performance. The best performance is achieved with the combination of ASR outputs from A, B, D, E and F, which is our submission system. The breakdown, over sessions, for the submission system are presented in Table 12. System C is excluded from the combination for the submission system as it uses the same AM as system D.

Table 11: *Performance of system combination on the development set.*

Systems combined	System	WER
W + U[1-6] - FBANK	A	63.0
W + U[1-6] - STE	B	62.8
A + B		62.0
A + B + D + E + F		**60.8**

Table 12: *Results for the submission system: WER (%) per session and location together with the overall WER.*

Test Set	Session	Kitchen	Dining	Living	Overall
Dev	S02	70.3	59.7	53.6	60.8
	S09	60.9	64.4	57.6	
Eval	S01	69.7	50.2	65.8	**56.5**
	S21	59.2	47.1	54.5	

The submission system achieved a performance of 60.8% WER on the *dev* set and 56.5% WER on the *eval* set respectively.

9. Conclusion

The Toshiba systems explored various enhancements, front-ends, AM architectures and speaker adaptation using VTLN for the final submission system. The system achieved a performance of 60.8% WER on the *dev* set and 56.5% WER on the *eval* set respectively. The system was ranked 4^{th} in category A, that focuses on acoustic robustness.

10. 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 2018*, Hyderabad, India, Sep. 2018.

[2] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[3] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 3214–3218.

[4] 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 *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 2751–2755.

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

[6] N. Derak, P. J. Kenny, R. Derak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1533–1545, October 2014.

[9] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," in *Proc. INTERSPEECH 2017*, Stockholm, Sweden, Aug. 2017, pp. 3832–3836.

[10] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. on Audio, Speech, Lang. Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[11] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE ASRU 2015*, Scottsdale, AZ, USA, Dec. 2015, pp. 444–451.

[12] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellerman, "Making Machines understand Us in Reverberant Rooms," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind Speech Dereverberation with Multi-Channel Linear Prediction Based on Short Time Fourier Transform Representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[14] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellerman, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal of Advances in Signal Processing*, 2016.

[15] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, 2017.

[16] D. R. Sanand and S. Umesh, "VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 20, no. 5, pp. 1573–1584, 2012.