CHiME 2018 Workshop on Speech Processing in Everyday Environments
07 September 2018, Hyderabad, India

DA-IICT/IIITV System for the 5th CHiME 2018 Challenge

Ankur T. Patil[1], *Maddala V. Siva Krishna*[2], *Mehak Piplani*[2], *Pulikonda Aditya Sai*[2], *Hardik B. Sailor*[1], *Hemant A. Patil*[1]

[1]Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India.
[2]Indian Institute of Information Technology (IIIT), Vadodara, Gujarat, India.

ankur_patil@daiict.ac.in, 201551045@iiitvadodara.ac.in, 201551072@iiitvadodara.ac.in, 201551013@iiitvadodara.ac.in, sailor_hardik@daiict.ac.in, hemant_patil@daiict.ac.in

Abstract

This paper presents our work on end-to-end (E2E) system development for multiple array track in the CHiME-5 challenge 2018. In particular, we propose to use E2E Lattice Free Maximum Mutual Information (LF-MMI) for acoustic modeling. For front-end, Mel Frequency Spectral Coefficients (MFSC) and Power Normalized Spectral Coefficients (PNSC) features are used. We employ delay-and-sum beamformer for speech enhancement of training and development data. The Recurrent Neural Network Language Model (RNNLM) rescoring is also explored along with *3*-gram language model. Our E2E LF-MMI Time Delay Neural Network (TDNN) system performed better than the E2E system provided in the challenge with an absolute reduction of 10.95 % in WER. The final system combination further reduces the WER to 78.63 %. Hence, our proposed system combination captures complementary information due to various E2E systems trained on full training data, beamformed data and using MFSC and PNSC features, respectively.

1. Introduction

Among the applications of robust ASR, voice command in domestic environments has attracted much interest recently, due in particular to the release of Amazon Echo, Google Home, Microsoft Cortana and other devices targeting home automation and multimedia systems. The performance in the robust ASR is significantly improving to due to advancement in signal processing, machine learning (and deep learning), speech enhancement and natural language processing (NLP) techniques [1] and availability of speech corpora in real noisy environment. The CHiME challenges and corpora along with DICIT, Sweet-Home, and DIRHA corpora have been significantly contributed in this area of research as these corpora contains the noisy speech signals in domestic environment. The difficulty level of the recognition task goes on increasing from CHiME-1 to CHiME-5 challenge. CHiME speech separation and recognition challenges aims to draw together the source separation and speech recognition communities as the source separation problems are inherent in the distant speech recognition.

Novelty of the CHiME-1 challenge corpus compared to previous available corpus is that the utterance to be recognized were provided in continuous audio background rather than as pre-segmented utterances thus allowing the range of background techniques to be employed [2]. CHiME-1 challenge was a small vocabulary recognition task and have fixed room impulse response. These limitations are overcame in CHiME-2 challenge by introducing the speaker movement and extending the corpus for medium vocabulary task [3]. The CHiME-3 challenge has considered speech recognition on a multi-microphone tablet device being used in noisy everyday environments [4]. CHiME-4 challenge uses same dataset as CHiME-3. However, it increases the level of difficulty by constraining the number of microphones available for testing [5]. The CMiME-5 challenge considers the problem of distant multichannel conversational ASR in everyday home environments with realistic noisy scenarios [6]. In this paper, we discuss about the our system implementation for CHiME-5 challenge.

The conventional hybrid DNN-HMM system is considered as the state-of-the-art in the distant multichannel ASR. Recently, there is a surge of developing end-to-end (E2E) ASR systems that can directly learn a mapping from an observation sequence to a target symbol. Various approaches to design E2E systems have been proposed, namely, Connectionist Temporal Classification (CTC) [7], attention-based [8], and RNN-transducer approach [9]. CTC is pioneering approach in E2E speech recognition which uses sequence level objective function. The limitation of CTC is that it fail to model the inter-dependencies at the output if output sequence is longer than the input sequence. However, RNN transducer extends the CTC defining the distribution over the output sequences of all lengths, and by jointly modeling both input-output and output-output dependencies. Attention based model belongs to the family of encoder-decoder models where encoder network maps the variable length input sequence to fixed length output sequence, represented by bottleneck features. These models requires very large data for the ASR system training. However, till now E2E ASR system does not perform well compared to the hybrid DNN-HMM system. In [10], authors recently proposed E2E ASR using Lattice Free Maximum Mutual Information (LF-MMI) and showed that it achieved comparable results to the hybrid LF-MMI DNN-HMM system.

The objective of this paper is to show the significance of E2E LF-MMI acoustic modeling for the CHiME-5 challenge. Specifically, we applied beamforming in the training, development, and evaluation set. The Mel Frequency [1]Spectral Coefficients (MFSC) and Power Normalized Spectral Coefficients (PNSC) features were used for acoustic modeling [11]. We also compared the performance of *3*-gram and RNNLM rescoring. Our final system combination experiments significantly improved the performance in the CHiME-5 task.

2. E2E LF-MMI Multichannel ASR System

The E2E model aims to train the neural-network-based acoustic model in one stage, i.e., without relying on alignments, building trees, or performing prior estimation [12, 10]. Among the sev-

[1]Here, Spectral coefficients refers to filterbank coefficients.

10.21437/CHiME.2018-17

eral end-to-end approaches, we used recently proposed Lattice Free Maximum Mutual Information (LF-MMI) method [10]. The objective function in this approach is maximum mutual information (MMI) in the context of Hidden Markov Models (HMM). MMI is a discriminative objective function E which aims to maximize the probability of the reference transcription, while minimizing the probability of all other transcriptions [10]:

$$E = \sum_{n=1}^{N} log \frac{f_\lambda(x^{(n)}|\mathbb{H}_{w^{(u)}})}{f_\lambda(x^{(u)})}$$

where λ is the set of all HMM parameters, N is total number of training utterances, and x represents speech utterance with corresponding transcription w. $\mathbb{H}_{w^{(u)}}$ is all possible state sequences pertaining to transcription $w^{(u)}$. The denominator can be approximated as,

$$f_\lambda(x^{(n)}) = \sum_{w} f_\lambda(x^{(n)}|\mathbb{H}_w) \approx f_\lambda(x^{(n)}|\mathbb{H}_{den})$$

where $\mathbb{H}_w$ is called as denominator graph that includes all possible sequences of words. Whereas, $\mathbb{H}_{w^{(u)}}$ is called as numerator graph. Using traditional method, computation of denominator graph was time consuming. In LF-MMI approach, they adopted few techniques to perform the denominator computation on GPU hardware. In regular LF-MMI context-dependent modeling is performed using tied biphone or triphone, where tying is done according to context-dependency tree. This prerequisite is removed in E2E LF-MMI approach by using monophones or full biphones. Composite HMM is used in the numerator graph instead of a special acyclic graph as in regular LF-MMI training. So, there is no restriction on self-loops in composite HMM to provide more freedom for the neural network to learn the alignments. Phone language model is estimated for the denominator graph using the training transcription. Context-dependency is implemented as a trivial full biphone tree. The detailed discussion on E2E LF-MMI model is given in [10].

We used two DNN architectures, namely, Time-Delay Neural Networks (TDNN) [13], and Long Short-Term Memory (LSTM) along with TDNN (TDNN-LSTM) [14]. To model the sequential data, such as time series, speech, etc., RNN is the first choice. The most effective and popular sequence models are used in the practical applications called as gated RNNs which include the LSTM [15]. The LSTM model is based on introducing self-loops to produce the paths, where the gradient can flow for a longer duration. Using the gate controlled by the hidden unit, the time scale of integration can be changed dynamically [15]. Another DNN architecture which has been shown to be effective in modeling the long range temporal dependencies is the TDNN proposed in [13]. In TDNN, initial layers learn representations using narrow context whereas higher layers learn wider context [13]. TDNN is one of the best performing systems tested in the Kaldi toolkit for various ASR task. We also used TDNN-LSTM system which is recently proposed to get advantages of both TDNN and LSTM models [14].

3. Experimental Setup

3.1. Database and Challenge Tracks

The database consists of the recording of twenty separate dinner parties taking place in real homes. These parties have been divided into disjoint training, development, and evaluation sets. The training, development, and evaluation data consists of 40:33, 4:27, and 5:12 hours of corpus recorded in 16, 2, and 2 parties with 32, 8, and 8 speakers having 79980, 7440, and 11028 utterances, respectively. More details about the CHiME-5 database is provided in [6, 16].

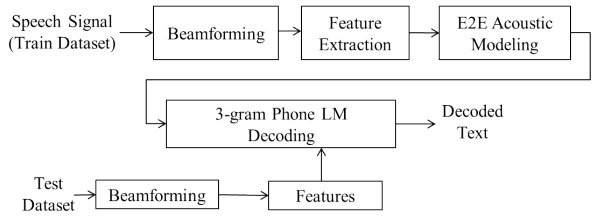

Figure 1: *Block diagram of LF-MMI E2E multichannel ASR system.*

The challenge features two tracks, namely, single array and multiple array. For each track, two rankings are produced. Ranking-A focuses on acoustic robustness only while ranking-B addresses all aspects of the task. We have developed system for multiple array which comes under ranking-B.

3.2. Pre-processing

The overall block diagram for the proposed approach is shown in Fig. 1. We have used the speech signals from all devices for given system development. Beamforming is performed on each device for training and development data [17]. We used delay-and-sum beamformer (BeamformIt toolkit [18]) applied on four microphone signals attached to reference array. It also reduces the size of training data by combining the signals from four channels to single enhanced signal. Features are extracted from each enhanced signal using a window length of 25 ms and shift of 10 ms. Feature extraction includes the use of Mel Frequency Spectral Coefficients (MFSC) and Power Normalized Spectral Coefficients (PNSC) with 40-D filterbank. We have also explored the Δ and $\Delta\Delta$ features.

3.3. Acoustic and Language Modeling

For acoustic modeling, we built LF-MMI based E2E system with TDNN and TDNN-LSTM architectures. The TDNN-based systems consist of 8 hidden layers with 2048 neurons per layer. The TDNN-LSTM-based systems consist of 7 TDNN and 3 LSTM layers with 1024 neurons per layer. The projection dimension of LSTM is 512 neurons. The L^2-regularization of 0.01 is applied in the hidden layers of both TDNN and TDNN-LSTM systems. For the softmax output layer, L^2-regularization of 0.0025 and 0.004 is used for TDNN and TDNN-LSTM systems, respectively.

The *3*-gram LM was built using the training text in the CHiME-5 challenge [19]. We have also built RNNLM using Kaldi-RNNLM toolkit which uses an importance-sampling based method to speed up training, and applies a new method to train unnormalized probabilities [20]. To build RNNLM, The embedding dimension of RNNLM is 1024. All the E2E ASR systems are trained in the Kaldi toolkit [21].

With possible combinations of speech enhancement technique, feature representations and DNN architectures for E2E system, we built several systems (S1-S5) using LF-MMI objective function. The specification of these systems is summarized in Table 1. The speech enhancement technique is applied to the training data for S2-S5 systems. We also used Δ and $\Delta\Delta$ features for these systems. The system S1 was built from the

speech signals from all channels and all devices training data without speech enhancement technique applied and using *40*-D MFSC features. The ASR system combination is performed using the Minimum Bayes Risk (MBR) technique with uniform weights to all the systems under consideration [22].

Table 1: *E2E LF-MMI ASR System Specification*

System	DNN Model	Features	Training Data
S1	TDNN	MFSC	Full Data
S2	TDNN	MFSC	Enhanced Speech
S3	TDNN-LSTM	MFSC	Enhanced Speech
S4	TDNN	PNSC	Enhanced Speech
S5	TDNN-LSTM	PNSC	Enhanced Speech

4. Experimental Results

The experimental results of the E2E LF-MMI TDNN and TDNN-LSTM systems using *3*-gram LM are reported in Table 2. The robust PNSC feature set did not perform well compared to the MFSC. The reason could be PNSCs features were performed well on simulated additive and convolutive noises [23]. The CHiME-5 database was recorded in real noisy environments along with the multiple speakers in conversation. The overall results show that the TDNN system performed well compared to the TDNN-LSTM for both the feature sets. The % WER in the kitchen environment is significantly higher compared to the living and dining hall as kitchen environment consists of multi-source noise. For all the systems, session S09 has a lower % WER compared to the session S02. We have also shown the experiments by combining several systems to get the complementary information. Out of many combinations, we reported the results for three combinations (denoted using $\oplus$ symbol), namely, SC-1 (S2 $\oplus$ S3), SC-2 (S1 $\oplus$ S2 $\oplus$ S3), SC-3 (S1 $\oplus$ S2 $\oplus$ S3 $\oplus$ S4). The best performance of 78.63 % WER is achieved by combining S1 to S4 systems (SC-3).

The experimental results of using RNNLM rescoring is reported in Table 3 for S1-S5 systems. The RNNLM rescoring improves the performance for S1-S3 systems with 0.14-0.35 % absolute reduction in WER. However, RNNLM rescoring did not improve the performance for S4 and S5 systems that used the PNSC feature set. In case of RNNLM rescoring, SC-2 combination gave the lowest % WER. The system combination experiments in Table 2 show that systems with *3*-gram LM significantly reduce % WER compared to the individual systems. However, combined systems (SC-1 to SC-3) did not improve the performance when using RNNLM rescoring compared to the *3*-gram LM. The reason may be presence of many disfluencies, irregular pronunciations, repeated words, and non-language tags in training corpus. For conversational ASR, RNNLM may not model them well. In such cases, combining RNNLM with explicit models, such as cache and trigger models [24], [25] or RNNLM adaptation [26] may improve the performance in conversational ASR.

In this challenge, we submitted the system results under multiple device track and ranking-B. The comparison of our proposed system with the baseline and other 3 systems submitted under the same track & ranking is shown in Table 4. The best performance is given by USTC-iFlytek system. The development of the best system includes iterative based speech separation, training data augmentation, SNR based array selection, and model fusions. The baseline system is conventional

Table 2: *Results of various E2E system and their combinations using 3-gram LM per session and location together with the overall % WER*

System	Session	Kitchen	Dining	Living	Overall
S1	S02	88.30	83.22	80.79	83.85
	S09	84.92	85.15	81.10	
S2	S02	88.62	83.01	80.54	83.75
	S09	84.42	85.51	80.88	
S3	S02	90.23	84.94	82.49	84.79
	S09	84.50	83.69	81.46	
S4	S02	89.13	85.65	83.93	85.17
	S09	84.82	83.95	81.97	
S5	S02	93.99	91.44	87.31	89.30
	S09	87.35	88.55	86.14	
SC-1	S02	85.89	79.88	77.49	80.14
	S09	79.79	79.44	77.06	
SC-2	S02	84.35	77.60	75.29	78.69
	S09	79.24	78.87	76.49	
SC-3	S02	84.21	78.46	75.64	**78.63**
	S09	78.23	78.15	76.27	

Table 3: *Results of various E2E system and their combinations using RNNLM rescoring per session and location together with the overall % WER.*

System	Session	Kitchen	Dining	Living	Overall
S1	S02	88.07	83.02	80.50	83.61
	S09	84.53	84.61	81.32	
S2	S02	88.56	83.09	80.10	83.40
	S09	83.38	84.68	80.94	
S3	S02	89.97	85.19	82.52	84.65
	S09	83.75	83.71	81.34	
S4	S02	89.50	86.65	84.21	85.64
	S09	85.04	84.75	82.47	
S5	S02	94.51	92.12	88.47	90.11
	S09	87.99	89.75	86.77	
SC-1	S02	86.47	80.43	77.97	80.64
	S09	79.69	80.41	77.76	
SC-2	S02	84.91	78.33	76.19	**79.04**
	S09	78.93	78.35	76.37	
SC-3	S02	84.79	79.27	76.54	79.36
	S09	79.12	78.68	76.83	

LF-MMI TDNN system which is trained with cleaned training data (removing irregular utterances) whereas our E2E LF-MMI system (S2) uses enhanced training data. Still, our E2E system (S2) performed significantly better than the ESPnet E2E baseline (10.95 % absolute reduction in WER). Finally, our proposed system combination gave an absolute reduction of 5.12 % in WER compared to the individual E2E system S2.

We have performed the experiment on evaluation set using our best system (SC-3) for challenge submission. The transcription is evaluated by challenge organizers. The results on evalu-

Table 4: *Comparison of proposed system combination with other systems on development system*

System	System Development						Session	Kitchen	Dining	Living	Overall
	BF	AU	E2E	RNNLM	MF	SA					
LF-MMI TDNN(Baseline) * [6]	√	-	-	-	-	-	S02 S09	87.3 81.6	79.5 80.6	79 77.6	81.3
ESPnet E2E ** [6]	√	-	√	√	-	-	S02 S09	- -	- -	- -	94.7
S2	√	-	√	√	-	-	S02 S09	88.62 84.42	83.01 85.51	80.54 80.88	83.75
Proposed Combination (SC-3)	√	-	√	√	√	-	S02 S09	84.21 78.23	78.46 78.15	75.64 76.27	**78.63**
USTC-iFlytec [27]	√	√	-	-	√	-	S02 S09	45.40 45.59	45.66 48.36	40.69 49.42	45.05
Hitachi-JHU [28]	√	√	-	√	√	√	S02 S09	59.31 50.64	52.96 50.69	48.95 50.46	52.38
CAS [29]	√	-	-	√	-	-	S02 S09	80.53 69.62	71.64 68.40	67.27 65.83	71.02

BF - Beamforming, AU - Data Augmentation, MF - Model Fusion, SA - Speaker Adaptation
* This system used data cleanup strategy [6].
** This system is built from ESPnet [30].

ation set using baseline system, best submitted system and our proposed system is reported in Table 5.

Table 5: *Results of multiple-array and ranking-B systems on evaluation set with the overall % WER.*

System	Session	Kitchen	Dining	Living	Overall
Baseline [6]	S01 S21	82.55 77.62	67.17 65.82	81.60 70.40	73.27
USTC-iFlytek [27]	S01 S21	58.08 52.47	37.11 41.11	55.07 42.20	46.14
Hitachi-JHU [28]	S01 S21	57.01 51.59	41.22 42.17	60.67 43.82	48.24
CAS [29]	S01 S21	67.72 65.13	53.61 53.14	72.91 58.45	61.01
Our Proposed System	S01 S21	82.65 79.49	73.38 72.55	84.68 69.82	76.42

5. Summary and Conclusions

In this study, we proposed to use E2E LF-MMI acoustic modeling for the CHiME-5 challenge. The beamforming technique is applied to the training and development set. This enhances the speech signals and also significantly reduces the training corpus in DNN training on single TITAN X GPU (7 days on full training database *vs.* 3 days on beamformed training database). We have also shown the impact of using RNNLM rescoring along with *3*-gram LM. It is observed that the *3*-gram LM performs slightly better than the RNNLM. It may be because of conversational speech and presence of many short utterances in the corpus. Our proposed framework of using E2E LF-MMI system performs significantly better than the challenge E2E baseline on development set. The system combination of TDNN and TDNN-LSTM trained from MFSC and PNSC features further improves the performance. We can modify the proposed system by training the E2E LF-MMI using BLSTM architecture. We can also train the individual ASR system for each device and fuse the systems by score combination technique. Recently proposed RNNLM adaptation technique can be employed for the task.

6. Acknowledgement

Authors would like to acknowledge the CHiME-5 challenge organizers for providing the corpus & offering travel grant to attend the workshop. We also acknowledge NVIDIA for the hardware grant of TITAN-X GPU. We are also thankful to authorities of DA-IICT for their kind support to carry out this research work.

7. References

[1] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 1–28, April 2018.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[3] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 126–130.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[5] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4^{th} CHiME speech separation and recognition challenge,"

URL:http://spandh.dcs.shef.ac.uk/chime_challenge { Last Accessed on 1 August, 2018}.

[6] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *INTERSPEECH 2018*, Hyderabad, India, Sep. 2018.

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, 2006, pp. 369–376.

[8] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in *NIPS, Workshop on Deep Learning*, Dec 2014, pp. 1–4.

[9] A. Graves, "Sequence transduction with recurrent neural networks," *ICML, Edinburgh, Scotland*, pp. 1–9, 2012.

[10] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *INTERSPEECH 2018*, Hyderabad, India, Sep. 2018.

[11] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, July 2016.

[12] 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, pp. 2751–2755.

[13] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH, Dresden, Germany*, 2015, pp. 2440–2444.

[14] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, March 2018.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. 1^{st} *Edition*. The MIT Press, 2016.

[16] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The 5^{th} CHiME speech separation and recognition challenge," URL:http://spandh.dcs.shef.ac.uk/chime_challenge { Last Accessed on 1 August, 2018}.

[17] M. Wölfel and J. McDonough, *Distant Speech Recognition*. John Wiley & Sons, 2009.

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.

[19] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *International Conference on Spoken Language Processing (ICSLP),Colorado, USA*, 2002, pp. 901–904.

[20] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 2018.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and J. Silovsky, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Big Island, Hawaii, USA*, 2011, pp. 1–4.

[22] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.

[23] Z.-Q. Wang and D. Wang, "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition," in *INTERSPEECH, Dresden, Germany*, Sep. 2015, pp. 2839–2843.

[24] R. Kuhn and R. D. Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, June 1990.

[25] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: A maximum entropy approach," in *ICASSP*, April 1993, pp. 45–48.

[26] K. Li, H. Xu, Y. Wang, D. Povey, and S. Khudanpur, "Recurrent neural network language model adaptation for conversational speech recognition," in *INTERSPEECH, Hyderabad*, Sept. 2018, pp. 1–5.

[27] J. Du, T. Gao, L. Sun, F. Ma, Y. Fang, D.-Y. Liu, Q. Zhang, X. Zhang, H.-K. Wang, J. Pan, J.-Q. Gao, C.-H. Lee, and J.-D. Chen, "The USTC-iFlytek systems for CHiME-5 challenge," in *CHiME-5 workshop*, Hyderabad, India, Sept. 2018.

[28] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N. E. Y. Soplin, M. Maciejewski, S.-J. Chen, A. S. Subramanian, R. Li, Z. Wang, J. Naradowsky, L. P. Garcia-Perera, and G. Sell, "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays," in *CHiME-5 workshop*, Hyderabad, India, Sept. 2018.

[29] J. Li, S. Xu, T. Wang, and B. Xu, "The ZTSpeech system for chime-5 challenge: A far-field speech recognition system with front-end and robust back-end," in *CHiME-5 workshop*, Hyderabad, India, Sept. 2018.

[30] ESPnet, URL:https://github.com/espnet/espnet, {Last Accessed: 25 July 2018}.