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, 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

The 5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018).

Evolution of CHiME Challenges

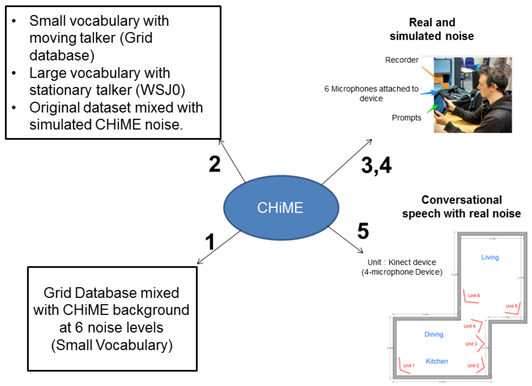

Figure 1: Evolution of database from CHiME-1 to CHiME-5.

- Top CHiME-1 and CHiME-2 systems came close to human performance.
- CHiME-3 best system : 5.8%WER
- CHiME-4 best system :2.2%WER
- CHiME-5 : The task of distant multi-microphone conversational speech recognition.
- Database of CHiME-5 Challenge [1]:
  - The scenario
    - Recording of twenty separate dinner parties taking place in real homes. (natural conversational speech).
    - Each party is of minimum 2 hours composed of 3 phases, corresponding to different locations namely, kitchen, dining, and living.
    - Natural movement from one location to other location is allowed.
  - Audio
    - 6 Kinect devices are strategically placed for each party.
    - Each Kinect device has linear array of 4 microphone.
    - Binaural microphone for each participant.

Table 1: CHiME-5 dataset

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

- Observations about database
  - Multiple speakers speaking at the same time.
  - Multisource noise is present.
  - Training text is not in context as it is conversational speech.

Block Diagram of Proposed ASR System

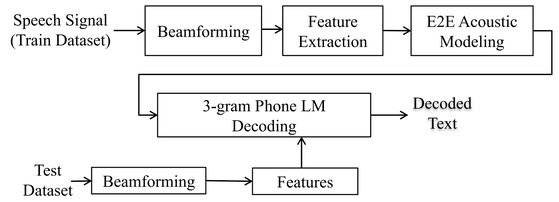

Figure 2: The comparison of ConvRBM filterbank scales with standard auditory frequency scales.

Proposed System Architecture Details

- E2E using LF-MMI [2] :
  - E2E : No prior alignments required from initial models, hence called flatstart.
  - Context dependency trees are not required.
  - Monophones or full left biphones are used for language modeling.
  - Entire training process in one stage.
- Speech enhancement : delay-and-sum beamformer. Applied on training and testing data [3].
- Feature extraction : Mel-frequency spectral coefficient (MFSC) and Power normalised spectral coefficient (PNSC) [4].
- Decoding is performed using 3-gram LM followed by RNNLM.
- Acoustic modeling using DNN:
  - Time Delayed Neural Network (TDNN): 8 hidden layers (2048 neurons per layer) [5].
  - Long-Short Term Memory along with TDNN (TDNN-LSTM) : 7-TDNN layers and 3-LSTM layers (1024 neurons per layer) [6].

System Specifications

Table 2: 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

* Full data : Audio data from all microphone channels and from all devices are used for training.
** Enhanced Speech : Audio data from each device is applied to beamformer. This enhanced speech is used for training.

- S1 : MFSC is 40-D (filterbank coefficients)
- S2-S5 : MFSC is 120-D (filterbank coefficients + Δ + $\Delta\Delta$)

Results-1

Table 3: Results of various E2E system and their combinations [7] using *3*-gram LM per session and location together with the overall % WER on development set.

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

- System combinations:
  - SC-1 : S2 $\oplus$ S3
  - SC-2 : S1 $\oplus$ S2 $\oplus$ S3)
  - SC-3 : S1 $\oplus$ S2 $\oplus$ S3 $\oplus$ S4

Table 4: Results of various E2E system and their combinations using RNNLM per session and location together with the overall % WER on development set.

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

Table 5: Results of our best system (SC-3) per session and location together with the overall %WER on evaluation set.

System	Session	Kitchen	Dining	Living	Overall
SC-3	S01 S21	82.65 79.49	73.38 72.55	84.68 69.82	76.42

Result-2

Table 6: Comparison of proposed system combination with baseline systems on development set.

System	Session	Kitchen	Dining	Living	Overall
LF-MMI TDNN*	S02 S09	87.3 81.6	79.5 80.6	79 77.6	81.3
ESPnet E2E **	S02 S09	- -	- -	- -	94.7
S2	S02 S09	88.62 84.42	83.01 85.51	80.54 80.88	83.75
SC-3	S02 S09	84.21 78.23	78.46 78.15	75.64 76.27	**78.63**

Summary and Conclusions

- Developed E2E system using LF-MMI as objective function.
- Performance in Kitchen is very poor due to presence of more multisource noise.
- RNNLM rescoring do not show any improvement because of conversational speech.

Acknowledgments

- CHiME-5 challenge organizers for providing database.
- NVIDIA for the hardware grant of TITAN X GPU.

References

1. 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.
2. 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.
3. M. Wolfel and J. McDonough, Distant Speech Recognition. John Wiley & Sons, 2009.
4. 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.
5. 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.
6. 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.
7. 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.