The JHU Multi-Microphone Multi-Speaker ASR System for the CHiME-6 Challenge

*Ashish Arora, *Desh Raj, *Aswin Shanmugam Subramanian, *Ke Li, Bar Ben-Yair, Matthew Maciejewski, Piotr Żelasko, Paola Garcia, Shinji Watanabe, Sanjeev Khudanpur

Center for Language and Speech Processing, Johns Hopkins University

draj@cs.jhu.edu

Abstract

This paper summarizes the JHU team's efforts in tracks 1 and 2 of the CHiME-6 challenge for distant multi-microphone conversational speech diarization and recognition in everyday home environments. We explore multi-array processing techniques at each stage of the pipeline, such as multi-array GSS for enhancement, posterior fusion for speech activity detection, PLDA score fusion for diarization, and lattice combination for ASR. We also integrate other techniques such as online multi-channel WPE dereverberation and VB-HMM based overlap assignment to deal with challenges like background noise and overlapping speakers, respectively. As a result of these efforts, our best system achieves a WER of 40.47% and 67.51% on tracks 1 and 2, respectively, on the evaluation set, which is an improvement of 10.82% and 10.43% over the baseline system for the challenge.

Index Terms: CHiME-6 challenge, robust speech recognition, speaker diarization, multi-channel, multi-speaker

1. System Description

Fig. 1 shows the components of our system for track 2. We use a similar system for track 1, except that the SAD and diarization blocks are replaced with the oracle segmentation.

1.1. Enhancement

1.1.1. Dereverberation and beamforming

We use an online version of the publicly available NARA-WPE [1] implementation of weighted prediction error based dereverberation for multi-channel signals [2] for all the channels in each array. This is followed by array-level beamforming using the BeamformIt tool [3]. All further processing is done on the beamformed signals.

1.1.2. Guided source separation (GSS)

Multi-array GSS [4, 5] is applied to enhance target speaker speech signals given the oracle speech segmentation in track 1 and the segmentation estimated via our speaker diarization in track2, respectively. We investigated the effectiveness of GSS by focusing on the neighboring context and show significant improvement (8.96% and 4.30%, respectively in tracks 1 and 2) with 20 second context.

1.2. Speech Activity Detection

We use a TDNN-Stats neural network trained to classify frames as $C = \{silence, speech, garbage\}$. The training targets are generated using alignments obtained from a GMM-HMM system. We further apply posterior fusion over the output distribution from all the arrays, i.e., for a frame $t, \forall i \in C, p_t(i) = f(p_t^k(i))$, where k denotes the arrays and f is the fusion criterion. We found that setting $f(\{x_i\}) = \max(x_i)$ is a simple and effective fusion scheme. We apply 0.01s frame shift and 0.2s padding on the speech segments during post-processing.

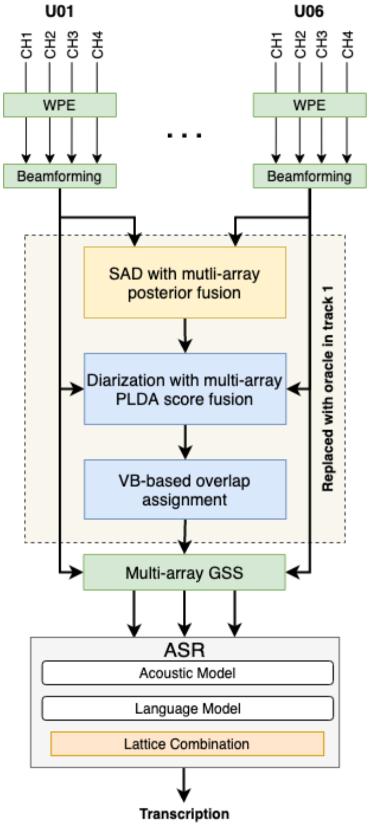
U01
CH1
CH2
CH3
CH4
WPE
Beamforming
U06
CH1
CH2
CH3
CH4
WPE
Beamforming
...
SAD with mutli-array posterior fusion
Diarization with multi-array PLDA score fusion
VB-based overlap assignment
Replaced with oracle in track 1
Multi-array GSS
ASR
Acoustic Model
Language Model
Lattice Combination
Transcription

Figure 1: *Overview of the decoding pipeline for track 2.*

1.3. Speaker Diarization

For diarization, we perform agglomerative hierarchical clustering (AHC) on the PLDA similarity scores computed between x-vectors [6] extracted from 1.5s windows with 0.25s shift [7]. The x-vector extractor consists of TDNN layers with stats pooling (similar to [8]), and is trained on VoxCeleb data [9] augmented with CHiME-6 background noises and simulated RIRs [10]. The PLDA parameters were trained on a 100k subset of the CHiME-6 training data. Similar to our score fusion method in SAD, we perform multi-array PLDA score fusion be-

[*] equal contribution

Table 1: *Speech activity detection results in track 2.*

System	Dev			Eval		
	MS	FA	Total	MS	FA	Total
Baseline (U06)	2.7	0.6	3.3	4.4	1.5	5.9
Fusion	1.1	0.8	1.9	2.4	2.8	5.2

Table 2: *Diarization results for track 2.*

System	Dev		Eval	
	DER	JER	DER	JER
Baseline (U06)	63.42	70.83	68.20	72.54
PLDA Fusion	63.97	71.65	71.56	71.32
+ 0.25s shift	61.00	66.23	69.64	69.81
+ overlap assign.	58.18	59.92	69.92	65.64

fore the clustering stage. Again, using $f(\{x_i\}) = \max(x_i)$ for the fusion function is found to be effective.

1.3.1. VB-based overlap assignment

Since AHC is not designed to handle overlapping speaker segments, we use a Variational Bayes Hidden Markov Model (VB-HMM) which leverages an LSTM-based overlap detector to assign frames to multiple speakers if an overlap is detected [11]. Finally as a post-processing step, we remove segments shorter than 200 ms.

1.4. Speech Recognition

1.4.1. Acoustic modeling

We incrementally train a monophone, a triphone (tri1), an LDA-MLLT (tri2), and a SAT (tri3) HMM-GMM systems, and use the tri3 model for generating alignments for neural network training. Before tri3 training, we also re-compute the pronunciation and silence probabilities using the tri2 system.

We train a CNN-TDNN-F model [12] on a combination of worn mic utterances (80h), beamformed array data (160h), and multi-array GSS enhanced data (40h) [13]. Cleanup is performed on this training data [14], which reduces it to $\sim$200h. We augment the resulting data with three-fold speed perturbation, but no reverberation is used. At inference time, we perform a 2-stage decoding similar to [15], where utterance-level i-vectors are used in the first pass, and then reweighted only with high confidence regions in the second pass.

1.4.2. Language modeling and rescoring

We use the transcription of training data to build our language models (LM). We use a 3-gram LM trained with the SRILM toolkit in the first pass decoding. For neural LMs, we use Kaldi to train recurrent neural network LMs (RNNLMs) [16]. We perform a pruned lattice rescoring [17] with a forward and a backward (reversing the text at the sentence level) LSTM. We first rescore with the forward LSTM and then perform another rescoring on top of the rescored lattices using the backward LSTM. Both LSTMs are a 2-layer projected LSTM model.

1.4.3. Lattice combination

For track 1, we use lattice combination method to combine three CNN-TDNN-F acoustic models. For track 2, we perform GSS with different input signals (multi-array beamforming and also individual array beamforming), and then we perform array level lattice combination. We decoded each array data using a common RTTM and on these decoded output, performed lattice combination.

Table 3: *WERs on Track 1.*

System	Dev	Eval
Baseline (U06)	51.75	51.29
CNN-TDNN-F AM	49.59	49.37
+Augmentation	44.51	44.92
+Overlap Feature	44.37	44.54
+LM Rescoring	42.82	42.94
+Lattice combination	41.75	42.07
+Lattice combination and Rescoring	40.27	40.47

Table 4: *WERs on Track 2.*

System	Dev	Eval
Baseline (U06)	84.25	77.94
CNN-TDNN-F AM	82.53	75.83
+PLDA Fusion	78.24	73.55
+GSS	70.97	68.75
+VB Overlap Assignment	69.28	68.78
+LM Rescoring	68.69	67.94
+Lattice combination	68.29	68.26
+Lattice combination and Rescoring	67.76	67.51

2. Challenge Results

2.1. Speech activity detection

Table 1 shows the results (missed speech (MS) and false alarm (FA)) for the posterior fusion method compared with the baseline. We used the original RTTM (obtained from transcriptions) for scoring the SAD.

2.2. Speaker diarization

We show results ablation for diarization in Table 2, scored using the force-aligned reference RTTM [1].

2.3. Speech recognition

We show improvement in WER from different modifications in track 1, Table 3. Our acoustic model consist of 6 convolutional layers and 16 TDNN-F layers trained with LF-MMI objective function. To train the model in similar to test conditions, we augmented the training data with test enhancements [13]. We added 1 bit oracle overlap information in neural network training which gave small improvement. The WER improvements from the frontend, our diarization, RNNLM rescoring and lattice combination are shown in Table 4.

[1] The DER is 5-6% better when scored using the original RTTM obtained from the reference transcriptions. We attribute this difference primarily to an increase in false alarms due to removal of short silence between words in the new RTTM.

3. References

[1] 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 *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[2] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[3] 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, 2007.

[4] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[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," *arXiv preprint arXiv:1905.12230*, 2019.

[6] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge." in *Interspeech*, vol. 2018, 2018, pp. 2808–2812.

[7] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, O. Plchot, O. Novotnỳ, H. Zeinali *et al.*, "But system description for dihard speech diarization challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[10] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[11] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection," *arXiv preprint arXiv:1910.11646*, 2019.

[12] 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 *Interspeech*, 2018, pp. 3743–3747.

[13] C. Zorila, 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," *arXiv preprint arXiv:1909.12208*, 2019.

[14] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6665–6669.

[15] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 539–546.

[16] 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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6109–6113.

[17] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5929–5933.