BUT System for CHiME-6 Challenge

Katerina Zmolikova, Martin Kocour[⋆]*, Federico Landini*[⋆]*, Karel Beneš*[⋆]*, Martin Karafiát*[⋆]*,
Hari Krishna Vydana, Alicia Lozano-Diez, Oldřich Plchot, Murali Karthick Baskar,
Ján Švec, Ladislav Mošner, Vladimir Malenovský, Lukáš Burget, Bolaji Yusuf,
Ondřej Novotný, František Grézl, Igor Szöke, Jan "Honza" Černocký*

Brno University of Technology, Faculty of Information Technology, IT4I Centre of Excellence, Czechia

izmolikova@fit.vutbr.cz

Abstract

This paper describes BUT's efforts in the development of the system for the CHiME-6 challenge with far-field dinner party recordings [1]. Our experiments are on both diarization and speech recognition parts of the system. For diarization, we employ the VBx framework which uses Bayesian hidden Markov model with eigenvoice priors on x-vectors. For acoustic modeling, we explore using different subsets of data for training, different neural network architectures, discriminative training, more robust i-vectors, and semi-supervised training on VoxCeleb data. Besides, we perform experiments with a neural network-based language model, exploring how to overcome the small size of the text corpus and incorporate across-segment context. When fusing our best systems, we achieve 41.21 % / 42.55 % WER on Track 1, for development and evaluation respectively, and 55.15% / 69.04 % on Track 2, for development and evaluation respectively.

1. Contributions

1.1. Diarization

For Track 2, we based our diarization on agglomerative hierarchical clustering (AHC) of x-vectors, followed by another x-vector clustering based on Bayesian hidden Markov model and variational Bayes inference (VBx). This approach was successfully applied to the Second DIHARD Challenge [2, 3, 4] and we adapted it to comply with the rules of the CHiME-6 challenge. We used the speech activity detection (SAD), x-vector extractor and probabilistic linear discriminant analysis (PLDA) modules from the baseline recipe [5, 6]. Moreover, we used only the enhanced recording from kinect U06 as in the baseline recipe. However, we extracted the x-vectors every 0.25 s instead of 0.75 s for having seen improvements previously [4].

Due to the high number of x-vectors to cluster, we used a two-step AHC: we divided the x-vectors from the whole recording into smaller groups accounting for several minutes, clustered each one individually and finally performed clustering with all the clusters to obtain four final clusters.

With a similar configuration but using a final threshold that allows for underclustering, we obtained the AHC-based initialization for VBx which was in turn run until convergence [4].

We further refined the diarization by a three-step procedure: First, we provided the speaker labels from the first VBx diarization to guided source separation (GSS) [7] to obtain four recordings where each speaker is enhanced in one of them. Then we ran the VBx diarization on each of the enhanced recordings and finally, we pooled the labels corresponding to the respective enhanced speakers to produce the final diarization.

The Performance of the methods is summarized in Table 1. We think the difference in DER trends between development and evaluation sets is explained in part by a larger proportion of silence in the latter and worse performance of the SAD system, which influences VBx performance more than AHC.

Table 1: *Comparison of different diarization methods in terms of Diarization Error Rate (DER) and Jaccard error rate (JER).*

	Development		Evaluation	
	DER	JER	DER	JER
Baseline	63.42	70.83	68.20	72.54
2-step AHC	60.21	65.21	71.84	71.80
VBx	51.67	53.20	75.11	71.77
VBx on GSS[1]	51.44	48.45	80.57	66.33

1.2. Enhancement

For the speech enhancement module, we have used the GSS method provided by the baseline. We applied the enhancement also on training data, using the oracle segmentation, and added the data to our training set for both Track 1 and Track 2 (referred to as *enhanced* in Section 1.3). In Track 2, we used the segmentation estimated by VBx diarization as guidance for GSS. For estimation of the masks, we used 40 seconds context on each side and the beamforming filters were estimated and applied every 5 seconds. Table 2 summarizes how the diarization and enhancement effect the automatic speech recognition (ASR) performance in Track 2. Results are obtained with the system corresponding to the first row in Table 4. Note that there is a slight inconsistency between these results caused by a different setting of decoding.

Table 2: *Impact of diarization and enhancement on the ASR performance on development set in Track 2. Column Diarization refers to the diarization used for ASR. We always used VBx diarization for guidance in GSS.*

Diarization	Enhancement	WER [%]
VBx	BeamformIt	73.4
VBx	GSS	62.9
VBx on GSS	GSS	59.6

[⋆] Equal contribution.

[1] In S01 we obtained five speakers so we slightly decreased VBx's F_A for that recording since we know that would allow for less speakers.

1.3. Acoustic model

We explored the impact of different subsets of training data on the performance of the acoustic model. The first combination consisted of the left microphone from worn data with all far-field data enhanced by GSS (*Worn (L) + enhanced*). Secondly, we used both worn microphones with enhanced data as in [8] (*Worn (S) + enhanced*). We further enlarged this dataset by adding worn data augmented with artificial room impulse responses [9] (+ *WornRVB*). The final combination used worn data from both microphones, enhanced data by GSS and a subset of 250k non-overlapped parts of far-field data (*+250k non-overlapped*). With all subsets, we applied speed perturbation and data cleaning as in the baseline. Table 3 compares the results of these combinations. In all cases the acoustic model was based on a convolutional time-delay neural network with semi-orthogonal factorization (CNN-TDNNf) [10]. In all experiments in this section, we used the baseline enhancement for Track 1 and for Track 2 we used GSS enhancement and VBx on GSS diarization.

Table 3: *Comparison of WER [%] on the development set when training the acoustic model with different data.*

		Track 1	Track 2
1	Worn (L) + enhanced	48.94	-
	(1) + w/o cleaning	49.14	-
2	Worn (S) + enhanced	47.85	59.29
3	(2) + WornRVB	47.57	59.22
4	(3) + 250k non-overlapped	47.31	59.02

We also explored other approaches for improving the performance of the acoustic model which we present in Table 4. These are: Basic CNN-TDNNf system (1), Re-transcription of the training data by basic CNN-TDNNf system and system retraining on pruned (beam 3) output lattices (2), and extension of lattice-free maximum mutual information with state-level minimum Bayes risk (sMBR) (3). We also considered incorporating segments from VoxCeleb [11, 12] data[2] shorter than 5 seconds for semi-supervised training (SST) (4). Then, we replaced the baseline offline i-vector extraction by the online version. In doing so, we dropped the baseline pseudo-speakers[3] and treated each person as a single speaker (5). Finally, we estimated a secondary offline i-vectors stream only on non-overlapped parts to help the system to do speaker separation (6).

Table 4: *Improvements of acoustic model using discriminative training, semi-supervised training on VoxCeleb and 2-stream i-vectors, compared on WER [%] on development set.*

		Track1	Track2
5	CNN-TDNNf	47.85	59.29
6	(5) + full-lattices	47.54	59.23
7	(5) + sMBR	46.37	58.86
8	(7) + VoxCeleb	45.81	57.25
9	(7) + speaker + online i-vector	45.36	57.12
10	(7) + non-overlapped + online i-vector	45.34	-

1.4. Language model

To improve over the baseline count-based language model, we have trained a long short-term memory language model (LSTM-LM) using BrnoLM toolkit[4].

To overcome the tiny size of the available data, we have combined two regularization techniques during training: the standard dropout along non-recurrent connections and random replacement of input tokens. We have obtained the best results with dropout 0.5 and input corruption rate 0.3.

We rescored 3000-best hypotheses from the ASR lattices, carrying over the hidden state between segments of each speaker. Table 5 shows the improvements achieved with model *Worn (S) + enhanced + WornRVB + 250k non-overlapped* from Table 3[5]. The gains were similar with other acoustic models.

Table 5: *Results of rescoring the development set of Track 1 with LSTM-LM consisting of two 650-units layers. Perplexity is for the each LM separately, WER is in interpolation of the respective LSTM-LM with the baseline.*

	Perplexity	WER [%]
baseline	157.7	48.24
+ LSTM	152.1	46.94
+ across-segment	136.5	46.61
+ input corruption	131.1	46.08

2. Final systems

For the final systems for both Track 1 and Track 2, we used ROVER [13] fusion over different acoustic models. In all systems we used GSS enhancement of test data, and for Track 2, VBx on GSS diarization.

For Track 1, we fused 8 systems: (3) + (4) + (5) + (6) + (7) + (9) + (10) (in Tables 3 and 4) + CNN-TDNNf system trained on *Worn (S) + enhanced + 250k non-overlapped* with full-lattices.

For Track 2, we fused 7 systems: (3) + (4) + (5) + (6) + (7) + (8) + (9) (in Tables 3 and 4). The final results in the submission format prescribed by the challenge are in Table 6 and 7.

Table 6: *Final results on Track 1, Categories A and B.*

	Development WER	Evaluation WER
Category A	42.75	44.34
Category B	41.21	42.55

Table 7: *Final results on Track 2, Categories A and B.*

	Development			Evaluation		
	DER	JER	WER	DER	JER	WER
Category A	51.44	48.45	55.60	80.57	66.33	69.17
Category B			55.15			69.04

[2] This data was allowed only for Track 2 but we analyzed this approach on Track 1 too.

[3] These were supposed to capture speaker movement.

[4] https://github.com/BUTSpeechFIT/BrnoLM

[5] The corresponding WER there is further improved by MBR decoding.

3. References

[1] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, and T. Yoshioka, "CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second DIHARD diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[3] F. Landini, S. Wang, M. Diez, L. Burget, P. Matějka, K. Žmolíková, L. Mošner, A. Silnova, O. Plchot, O. Novotný, H. Zeinali, and J. Rohdin, "BUT System for DIHARD Speech Diarization Challenge 2019," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[4] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černocký, "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs." in *Interspeech*, 2016, pp. 3434–3438.

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

[7] 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.

[8] 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," 2019.

[9] 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)*, 2017, pp. 5220–5224.

[10] 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 *Proceedings of Interspeech*, 09 2018, pp. 3743–3747.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

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

[13] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, 1997, pp. 347–354.