Towards a speaker diarization system for the CHiME 2020 dinner party transcription

Christoph Boeddeker[1], Tobias Cord-Landwehr[1], Jens Heitkaemper[1], Cătălin Zorilă[2], Daichi Hayakawa[3], Mohan Li[2], Min Liu[4], Rama Doddipatla[2], Reinhold Haeb-Umbach[1]

[1]Paderborn University, Department of Communications Engineering, Paderborn, Germany
[2]Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom
[3] Toshiba Corporation Corporate R&D Center, Kawasaki, Japan [4] Toshiba China R&D Center, Beijing, China

boeddeker@nt.upb.de, catalin.zorila@crl.toshiba.co.uk

Abstract

In this work, we present our joint efforts on *Track 2* of the CHiME-6 challenge, where a transcription of a dinner party is to be done on 2 to 3 hour sessions without the use of start and end time annotations for each utterance during evaluation. The first contribution follows the challenge guidelines and combines our system presented during the last challenge [1] with the Track 2 baseline diarization system [2]. Different acoustic models (AMs) with system combination are tested on the enhanced data. The second contribution violates the challenge rules but allows an outlook on a system which may achieve strong results if the oracle component is replaced in the future.

Index Terms: speaker diarization, speech recognition, permutation invariant training

1. Guided Source Separation

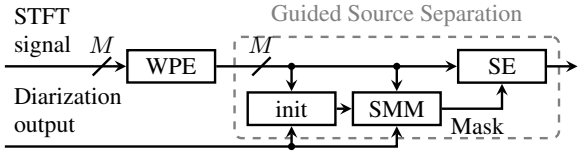

Figure 1: *Blockdiagram of the GSS system introduced during the CHiME-5 challenge.*

The enhancement system presented in [1] relies on the human annotations of start and end times for each speaker that are not allowed to be used during Track 2 of the CHiME-6 challenge. However, the system can still be applied under the new challenge rules if the annotation information is replaced with the output of a diarization system as shown in Fig. 1 [1]

2. Acoustic model

The baseline acoustic model (AM_0) provided by the challenge has a 15-layer factorized time delay neural network (TDNNF) topology and is trained using a combination of 40-dim MFCCs and 100-dim i-vectors. Training data is formed of unprocessed and artificially reverberated speech. We have tested five additional HMM-DNN AMs consisting of combinations of convolutional neural networks (CNNs) with or without residual connections and TDNNFs, as follows. AM_1, AM_2 and AM_3 use 10 CNN layers followed by 9 TDNNF layers, while AM_4 and AM_5 use 40 RESNET CNN layers. AM_1 is trained using unprocessed worn and WPE processed array data, AM_2 is trained using unprocessed worn and GSS processed array data, and AM_3 applies discriminative training (DT) on top AM_2. AM_4 is trained on the same data as AM_2 but uses a RESNET architecture, and AM_5 is based on AM_4 with DT. Both the baseline 3-gram language model (LM) provided with the challenge and an RNN LM comprising of 3 TDNN and 2 LSTM layers were used for scoring.

[1]After the challenge we will publish the modification for the GSS system on https://github.com/fgnt/pb_chime5.

Table 1: *DEV (EVAL) ASR results for Track 2 using baseline diarization system.*

Enh. in test	ASR	WER (%)	
		3G-LM	RNN-LM
WPE+BFIt	AM_0	81.92 (76.37)	-
GSS	AM_0	78.12 (73.06)	77.71 (72.47)
	AM_1	76.44 (72.04)	75.93 (70.80)
	AM_2	74.74 (71.27)	74.15 (70.42)
	AM_3	74.67 (70.55)	74.23 (70.07)
	AM_4	74.05 (70.47)	73.79 (70.05)
	AM_5	74.73 (70.14)	74.42 (69.64)
	AM_{0-5}	73.50 (68.96)	73.05 (68.45)

Table 2: *Submitted results*

	Development set			Evaluation set		
	DER	JER	WER	DER	JER	WER
Cat. A	62.61	70.95	73.50	66.93	71.44	68.96
Cat. B	62.61	70.95	73.05	66.93	71.44	68.45

3. Experiments

Results of ASR experiments using the baseline diarization system provided by the challenge are depicted in Table 1 for the development (DEV) and evaluation (EVAL) sets. The performance with in-house retrained version of AM_0 is also provided. Test data for AM_0 were processed using WPE (over the complete session) followed by BeamformIt. Applying GSS on the test data (with input from the diarization system instead of human annotations) yields a significant gain in accuracy as shown in Table 1. A further WER improvement is achieved by combining the lattices of systems 0 to 5, for both the 3G and RNN LMs.

4. PIT Neural Speaker Diarization

An alternative approach to the baseline diarization system [2] is to formulate the diarization problem as a multi-class labeling problem, as proposed in [3]. This is particularly interesting in the context of CHiME-6, because this Neural Speaker Diarization (NSD) naturally handles overlapped speech and because in CHiME-6 the total number of speakers (4) is fixed and known in advance. This single system replaces the speech activity detector (SAD) estimator, the speaker embedding calculation and the clustering of the baseline diarization system.

For each speaker the start and end times of an utterance (or word) have to be derived from the estimated speech presence probability. In [3] this was solved with a threshold and a median filter. Here, a simple Viterbi decoding on an HMM as used in the SAD baseline system is applied independently to the speech presence probability of each speaker to obtain the start and end times.

To train the Neural Network (NN) we use a permutation invariant training (PIT) objective [4, 5], i.e., compute the loss for each permutation of target speaker activity label and network output, and back propagate the minimum loss. In contrast to the original PIT loss the system is not dependent on parallel data since the targets only consist of activity information. The activity information is estimated using an acoustic model to calculate a forced alignment for each speaker and setting the speaker to active for all frames assigned to non-silence senones.

5. Spatial features

We trained the NSD system on the CHiME-6 data, but the diarization error rate (DER) on the training dataset stayed relative high, although the system was already overfitting to the training data. The reason is that there are only 32 speakers in the CHiME-6 training set, which is far too few to generalize well to unseen speakers in the test set.

To improve the performance and generalizability we investigated options to add spatial information to the system. It helps discriminating speakers, and since a spatial feature is not directly linked to a specific speaker it may also improve generalization for training data with a low number of speakers.

Spatial information has shown to improve the results for source separation on various databases [6, 7]. Common features are the inter-channel phase differences (IPD) [6]. However, the angles between the speakers relative to the array, i.e., their spatial resolution, is quite small and our preliminary experiments showed that even some spatial mixture models had problems utilizing the spatial information on this dataset [1]. For this reason, the use of IPD features was discarded.

A spatial mixture model (SMM) [8] is an unsupervised method for source separation and has shown to achieve strong results as part of the Guided Source Separation (GSS) system [1]. In this work we used the posterior probabilities of speaker presence obtained in the E-Step of the EM algorithm of the SMM training as additional input features to the NSD system. To be specific, we calculated the average power across all channels and frequencies of the observation weighted with the posterior mask to be used as spatial features.

One could argue that the speaker presence posteriors of the EM algorithm applied to the SMM are already the sought-after diarization information, but initial tests revealed that they were too noisy and unreliable.

6. Limitations and open problems

Ideally, the NSD system should operate on the complete session of 2 to 3 hours length to estimate the diarization information. This is difficult, first because the memory consumption it too high and, second, because the information about the past that can be stored in a recurrent network node is limited. Therefore, the session is first split in segments of fixed length. This, however, introduces a segment permutation problem, because the NN does not output the same speaker on the same index for every segment. In this work, we do not address this problem and use an oracle permutation solver instead. This is not in line with the challenge guidelines so that the NSD system will not be ranked. To ease the task of the (still to be developed) permutation solver we used relatively large segments of 40 seconds length.

7. Experiments

In Table 3 some experimental results with the NSD system are shown. As network architecture either one or two BLSTM layers are used, followed by two dense layers. The input are Mel features, on which VTLP [9] is applied during training. The aforementioned spatial features are concatenated with the Mel features. The output of the NSD system is used as input to the GSS [1] followed by the baseline Automatic Speech Recognition (ASR) system.

The first line in Table 3 is the baseline[2]. The first NSD system without spatial features reached $34.28\,\%$ DER on the training data and $60.09\,\%$ DER on DEV+EVAL. The results indicated that the model overfits to the training data without perfectly learning the targets. Including the spatial features from the SMM reduces the DER to $2.52\,\%$ on train and $53.75\,\%$ on DEV+EVAL. For the training data the spatial information allows the system to achieve close to perfect results, but the results on DEV+EVAL are far from perfect, asking for more techniques to reduce overfitting. The last line in the Table uses the diarization provided by the challenge organizers for Track 1. This result is included to assess how much performance is lost by the non-perfect diarization system.

Table 3: *Preview experiments with NSD supported by SMM spatial features. DER is averaged across DEV and EVAL. "Old" and "new" refers to the DER targets that changed during the challenge.*

Network	Mel bins	SMM spat. features	Dropout	DER			WER	
				TRAIN	DEV+EVAL old	DEV+EVAL new	DEV	EVAL
Baseline[2]	-	-	-	-	60.87	64.74	77.49	71.92
2 BLSTM	80	No	No	34.28	60.09	58.56	72.68	71.25
2 BLSTM	80	Yes	No	2.52	53.75	68.84	70.05	69.50
2 BLSTM	80	Yes	0.25	2.79	52.71	65.80	68.65	67.43
1 BLSTM	80	Yes	0.25	5.11	50.99	66.54	65.65	66.60
1 BLSTM	**24**	**Yes**	**0.25**	**36.4**	**47.59**	**57.00**	**62.50**	**64.59**
Oracle	-	-	-	0	0	47.67	49.54	

8. Acknowledgements

Computational resources were provided by the Paderborn Center for Parallel Computing.

[2]Note: Here, we used the same baseline system as in Table 1, but it was trained in a different environment. Therefore the numbers are different.

9. References

[1] 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.

[2] 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.

[3] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," *arXiv preprint arXiv:1909.05952*, 2019.

[4] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[6] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[7] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, 2019.

[8] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1153–1157.

[9] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.