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, cord, heitkaemper, haeb}@nt.upb.de
{catalin.zorila, mohan.li, rama.doddipatla}@crl.toshiba.co.uk
daichi1.hayakawa@toshiba.co.jp, liumin@toshiba.com.cn

Abstract

In this work, we present our joint efforts on *Track 2* of the CHiME 6 challenge, where two to three hours long sessions of a dinner party are to be transcribed without the use of start and end time annotations for each utterance during evaluation. The first contribution introduces an extension to an earlier proposed neural speaker diarization system by additionally incorporating spatial features, but violates the challenge rules by using oracle information about the speaker permutation in different segments. However, the results are promising and warrant future investigations. The second contribution follows the challenge guidelines and combines our system presented during the last challenge with the Track 2 baseline diarization system. Different acoustic models with system combination are tested on the enhanced data and deliver significant performance improvements over the baseline, both with the baseline and a modified language model.

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

1. Introduction

Automatic Speech Recognition (ASR) using neural networks has shown impressive results on multiple databases. However, most systems still underperform in the present of reverberation and overlapping speech. One example for a scenario with such adversarial conditions are dinner party or meeting scenarios. The CHiME 5 challenge [1] presented a database of real audio recordings with a high amount of overlapping speech, reverberation and noise. During and after the challenge, a number of contributions presented systems which were able to achieve large improvements in recognition accuracy over the baseline system even in these adverse conditions [2, 3, 4, 5, 6].

However, all of these approaches relied on the provided speaker activity information, which may not be available in real world applications. Therefore, recent publications have focused on developing diarization techniques to enable robust ASR on continuous speech databases without speaker activity information [7, 8]. Most of the efforts to improve diarization rely on neural networks and can be divided into two main groups. The first approach uses data with a large variety of speakers to train a speaker embedding extractor which is applied on multi talker data during testing to identify the active speakers [9, 10]. During testing, the signal is divided into multiple segments on which a speaker embedding is extracted and a clustering algorithm is used to assign a speaker to each segment. Here, a mismatch between the training and evaluation data can be mitigated through domain adaptation algorithms, data augmentation and sufficiently diverse training data [11]. The second approach trains a neural network based diarization system directly on the multi talker signals [12, 13]. During the DIHARD II challenge mainly diarization systems of the first group have been tested on the CHiME 5 database with varying results [14].

The second track of the CHiME 6 challenge disallowed the use of the oracle speaker activity information during testing [15]. Without the activity information no utterance boundaries are defined, so that the system input consists of two-hour sessions of speech by four speakers from changing positions in up to three rooms. Therefore, the baseline for the CHiME 6 challenge includes a diarization system, which follows the first approach mentioned above, to compensate for the lack of speaker activity information compared to the CHiME 5 challenge [15].

In this paper, we propose to use a Permutation Invariant Training (PIT)-Neural Speaker Diarization (NSD) system [12] as part of the front-end for an ASR system on the CHiME 6 database, which is inspired by the aforementioned second approach. Additionally, an adjustment of our Guided Source Separation (GSS) system proposed for the CHiME 5 challenge [4] is presented, which allows the use of the diarization output instead of the oracle speaker activity information.

The PIT-NSD is applied independently to segments of the two-hour session which may lead to a permutation of the speakers between segments. To solve this permutation one may use speaker embeddings extracted on single speaker chunks. However, in this work we concentrated on the NSD system and used oracle information about the speaker permutations, leaving the implementation of a permutation solver for future work.

To improve upon the accuracy of the diarization estimation a novel spatial feature relying on information gained from a spatial mixture model is presented.

Since the proposed diarization system uses an oracle permutation solver, it was not part of the challenge submission. For the submission, we adjusted the GSS front-end to accept the output of the baseline diarization system to show results conforming to the challenge regulations. Further experiments are done with different acoustic models and acoustic model combination to improve upon the recognition results.

The remainder of this paper is structured as follows. First, the CHiME 6 database is introduced in Section 2 and a short system overview is presented in Section 3. In Section 4 the proposed diarization and the novel spatial features are discussed and in Section 5 the modification to the GSS front-end are explained. Section 6 gives an overview over the back-end used in this work. We close the paper with some evaluation in Section 7 and a short conclusion.

2. Database

This section gives a short overview over the CHiME 6 data. The database consists of recordings of several dinner parties by 4 speakers, where each party consists of around 2 hours of audio. During the parties the speakers prepare food, dine and socialize. The three activities take place in different rooms and each lasts for at least 30 minutes. The signals were captured by six Microsoft Kinect microphone arrays with four audio channels each and two arrays per room. Additionally, each speaker is recorded by binaural in-ear microphones. However, the captured in-ear audio includes both overlapping speaker and noise so that no clean, parallel recording of the microphone array data is available to train a neural network-based source separation system in the usual supervised training setup. The training dataset comprises a total of 40 hours of audio from 32 speakers, while the development and evaluation set consist of five hours from 8 speakers each. All participants were encouraged to speak naturally, so that more than 20 % of the recorded time the speakers overlapped.

3. System Overview

A common approach to estimate the transcription for multiple speakers in an unsegmented, long audio stream (e.g. 2h long audio) is to utilise different system components to solve the task step by step [15]. The first component splits a session in active and inactive chunks for each speaker. Such component is called a diarization system. It estimates the start and end times of each utterance and speaker, where an utterance may range from a single word to one or more sentences. Depending on the architecture, a diarization system can either handle overlapping speech and identify start and end times of each involved speaker's speech or it ignores segments with concurrent speakers. The diarization is followed by a speech enhancement and/or extraction component. While the speech enhancement can in principle also be applied before the diarization system, it is more common that the speech extraction uses the diarization information to identify which speaker to extract. Finally, an ASR system takes the segments where a single speaker is active (and extracted) as input and predicts the transcription for the utterance.

4. Diarization

4.1. CHiME 6 baseline

The baseline diarization system for CHiME 6 contains the following components [15]. First, a Speech Activity Detector (SAD) Neural Network (NN) evaluates each frame for speech/voice activity. As a postprocessor, a simple Viterbi decoding on an HMM estimates the onsets and offsets of speech activity predicted by the network. The segments with active speech are then divided into smaller subsegments, on each of which another neural network extracts x-vectors, i.e., speaker embeddings. This is followed by Agglomerative Hierarchical Clustering (AHC) of the embedding vectors using a similarity measure derived from Probabilistic Linear Discriminant Analysis (PLDA), until the number of clusters matches the number of speakers (four in the CHiME scenario). The speaker labels for each cluster are then given by the cluster assignments.

The x-vector extractor itself has been trained on a large external database (VoxCeleb [16]). Note that the clustering naturally solves the problem of re-identifying a previously active speaker after he/she has been silent for some time. The main drawback of this approach is, however, that it cannot handle overlapping speech. If multiple talkers are active at a time, the system will at most identify one speaker correctly.

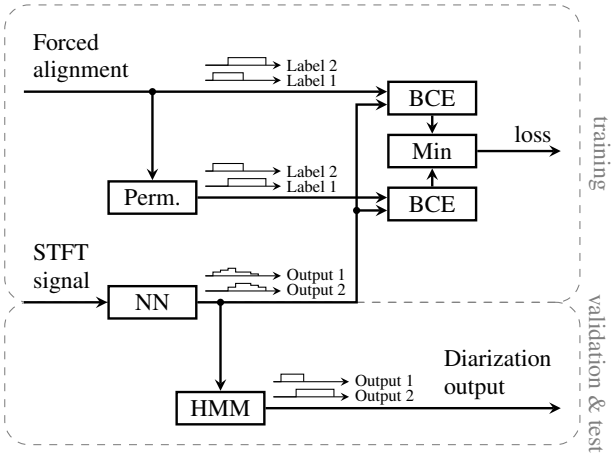

Figure 1: *Permutation invariant training (PIT) for a neural speaker diarization system. Inspired from [12, Fig. 1]. The forced alignment are produced from an ASR system. They mean here the refined human annotations [5].*

4.2. PIT Neural Speaker Diarization

An alternative approach to the diarization system in Section 4.1 is to formulate the diarization problem as a multi-class labelling problem, as proposed in [12] and visualized in Fig. 1. This is particularly interesting in the context of CHiME 6, because this 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 SAD estimator, the speaker embedding calculation and the clustering of the baseline diarization system.

To train the NN we use a PIT objective [17, 18], 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 used for training source separation systems, the system here does not require parallel data for training 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.

For each speaker the start and end times of an utterance (or word) have to be derived from the speech presence probability estimated by the neural network. In [12] 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 utterance start and end times.

4.3. Spatial features

The NSD system tends to overfit when the training data has a low number of speakers. In the CHiME 6 training set there are only 32 distinct speakers, which is far too few to generalize well to unseen speakers in the test set[1].

[1]For comparison, the baseline diarization system is trained on more that 7000 speakers from VoxCeleb

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 [19, 20]. Common features are the inter-channel phase differences (IPD) [19], 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 [4]. For this reason, the use of IPD features was discarded.

Here, we propose to use a Spatial Mixture Model (SMM), the complex angular central Gaussian mixture model (cACGMM) [21] to be specific, to obtain spatial features. This mixture model has shown to achieve strong results as part of the GSS system [4] when a guide (human diarization output) is available.

The parameters of a SMM are estimated with the EM algorithm. The output of the Expectation step, the posterior probabilities of each speaker being present in a certain time-frequency bin, can be used as masks to extract the speakers. However, in CHiME 6 we observed that the SMM has problems to reliably estimate the masks for all speakers[2]. So instead of using the masks to extract the speakers we propose to use the masks to extract simple features as additional input for 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.

5. Guided Source Separation

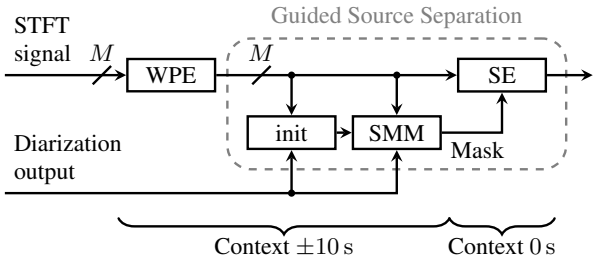

Figure 2: *Block diagram of the GSS system introduced during the CHiME-5 challenge.*

The enhancement system presented in [4] 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. 2 [3].

The first step in the enhancement system is dereverberation, which is performed with the Weighted Prediction Error (WPE) algorithm [22, 23, 24]. Source extraction (SE) is done with a Minimum Variance Distortionless Response (MVDR) beamformier in the formulation presented in [25], using the reference channel estimation as in [26]. Further the beamformer output is processed by blind analytic normalization (BAN) postfilter [27] to reduce speech distortions.

To compute the beamformer coefficients the spatial covariance matrices of speech and distortions have to be estimated. To this end, the SMM (i.e., cACGMM [21]) estimates posterior probability masks for each speaker and noise. It should be mentioned that the SMM uses the diarization output twice, firstly, as initialization of the Expectation Maximization (EM), and secondly, as a constraint for the posterior probability masks. Interestingly, the SMM relies on the diarization output as guide to improve the mask estimation, while the diarization system employs spatial features in turn computed with the help of a SMM.

In contrast to the original system in [4], the temporal context for WPE and the SMM parameter estimation (see Fig. 2) is reduced from $\pm 15\,\mathrm{s}$ to $\pm 10\,\mathrm{s}$ and the number of channels is reduced from 24 channels to 12 channels (only the outer microphones of each array are used). These changes were made to reduce the computational load, and have only small effects on the Word Error Rate (WER). A further change, compared to [4] is that the power spectral density matrix for the beamformer estimation is calculated on the masked observation without additional context, as proposed in [5].

6. Acoustic and language models

Three distinct acoustic models (AMs) architectures were used in our experiments. One was the baseline AM provided by the challenge organizers consisting of a 14-layer Factorized Time Delay Neural Network (TDNNF) topology. This model was trained with unprocessed and artificially reverberated speech using a combination of 40-dim MFCCs and 100-dim i-vectors. Second acoustic model was formed of 10 convolutional neural network (CNN) layers followed by 9 TDNNF layers (CNN-TDNNF), and the third model consisted of 40-layer CNNs with residual connections (RESNET).

The training set of the latter two models contained unprocessed worn and enhanced array data, augmented by 3-fold speed perturbation. The speech enhancement was either dereverberation (WPE) or GSS. The training was performed in KALDI using LF-MMI criterion [28], and the acoustic features were 64-dim FBANK combined with 100-dim i-vectors. Discriminative training (DT) was also applied on top of previously trained LF-MMI models.

Both the baseline 3-gram language model (LM) provided with the challenge and an RNN LM consisting of 3 TDNN and 2 LSTM layers were used for scoring. For more details about the AMs and the RNN LM see [29].

7. Experiments

This section is divided into two parts. The first part describes the results with the PIT-NSD diarization system and an oracle permutation solver using the baseline acoustic model. In the second part, the diarization system is replaced with the baseline diarization to assure challenge rule conformity. Additionally, multiple acoustic models are compared and combined to reduce the WER. All systems are compared in terms of both WER and DER. The DER is calculated either with non-silent alignments obtained by an HMM-GMM acoustic model or human annotations as ground truth. We present the DER for both types of ground truth activity information since the non-silent alignments are the official challenge ground truth whereas our experiments have shown a better correlation between the WER

[2]Note: The SMM in the GSS system [4] used the human annotations to tackle this problem

[3]We published the modification for the GSS system on https://github.com/fgnt/pb_chime5 .

Table 1: *Experiments with NSD supported by spatial features. Diarization Error Rate (DER) is averaged across DEV and EVAL. "Annotation" and "Alignment" refer to the DER targets that changed during the challenge. Oracle segment permutation information is assumed.*

Network	Mel bins	SMM spatial features	Dropout	DER			WER in %	
				TRAIN	DEV+EVAL		DEV	EVAL
					Annotation	Alignment		
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**	**7.03**	**48.89**	**64.65**	**63.82**	**63.47**
Oracle	-	-	-	-	0	38.16	47.67	49.54

and DER calculated from human annotations.

7.1. Limitations and open problems of NSD on CHiME 6

Ideally, the NSD system should operate on the complete session of 2 to 3 hours 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 into 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 the permutation problem between segments and use an oracle permutation solver instead. To ease the task of the (future) permutation solver we used relatively large segments of 40 seconds length. However, the spatial features would benefit from even longer segments since the spatial mixture model requires the prior knowledge of the number of active speaker which is set to four for all segments. Longer segments increase the likelihood that all speakers are active for at least some time during the segment. Additionally, the PIT-NSD tends to overfit to the training data in case of a low number of speakers during training, which is the case for the CHiME 6 data. Despite these problems, the NSD system has the potential to achieve strong diarization estimates without an external database for training a speaker embedding extractor, as will be shown in the next section.

7.2. NSD and spatial features

In Table 1 experimental results with the NSD system are shown. The network architecture is adapted from common u-PIT architectures [30] and consists of either one or two Bidirectional Long Short-Term Memory (BLSTM) layers followed by two dense layers.

The input are Mel features concatenated with the spatial features described in Section 4.3. During training, VTLP [31] is applied to the Mel features to reduce network overfitting.

The DER for the TRAIN data is calculated against the training target, which are the forced alignments of an ASR system, similar to the provided alignment for DEV and EVAL.

The first line in Table 1 represents the challenge baseline system. In the second line, the results for the NSD system without spatial features are shown which achieves $34.28\,\%$ DER on the training data and $60.09\,\%$ DER on DEV+EVAL. The results indicate that the model overfits to the training data as expected. 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. Therefore, both a reduction in parameters by using only one BLSTM layer or reducing the number of Mel features and the application of dropout are examined in further experiments. All these changes lead to an improvement in DER on both the DEV and EVAL datasets, while slightly increasing the DER on the TRAIN set, thus indicating a small reduction in overfitting. However, the gap between the diarization results on the TRAIN and EVAL/DEV set is still quite large, which asks for further improvements, e.g., applying more advanced data augmentation techniques to reduce overfitting. All WER results on the DEV and EVAL mirror the improvements observed for the DER. The last line in Table 1 uses the diarization provided by the challenge organizers for Track 1. This result is included to assess how much performance is lost by the real diarization system.

7.3. Challenge Contribution

The results shown in the previous section are not in line with the challenge rules. Therefore, the PIT-NSD system is replaced with the baseline diarization described in Section 4.1. In Table 2 the second line represents the result of this system with the baseline acoustic model. The GSS system already leads to an improvement of approximately $3\,\%$ compared to the baseline front-end for both DEV and EVAL data. Table 2 includes additional results with the acoustic models described in Section 6 and offers a comparison between an Recurrent Neural Network (RNN) based language model and the baseline 3-gram model. The RNN language model outperforms the 3-gram model in all cases. The last line presents the results combining the lattices of all six proposed acoustic models. Both the results with the baseline and the RNN-based language model achieve the fourth best WER in their respective category during the CHiME 6 challenge with $68.96\,\%$ and $68.45\,\%$ WER. In Table 3 the results submitted to the challenge are displayed.

Table 2: *DEV and EVAL ASR results for Track 2 using baseline diarization system.*

Enh. in test	ASR			WER in %			
	Training Data	Acoustic model		3G-LM		RNN-LM	
		Topology	DT	DEV	EVAL	DEV	EVAL
WPE+BFIt	unproc. & reverb.	TDNNF(14)		81.92	76.37	-	-
GSS	unproc. & reverb.	TDNNF(14)		78.12	73.06	77.71	72.47
	worn & WPE	CNN-TDNNF(19)		76.44	72.04	75.93	70.80
	worn & GSS	CNN-TDNNF(19)		74.74	71.27	74.15	70.42
	worn & GSS	CNN-TDNNF(19)	✓	74.67	70.55	74.23	70.07
	worn & GSS	RESNET(40)		74.05	70.47	73.79	70.05
	worn & GSS	RESNET(40)	✓	74.73	70.14	74.42	69.64
		Latt.Comb.		73.50	68.96	73.05	68.45

8. Conclusion

In this paper we presented a neural speaker diarization system for the CHiME 6 database. Our first contribution was using the posterior masks from a spatial mixture model as features. The additional features lead to a large improvement in DER on the training set but only to a smaller reduction in DER on the DEV and EVAL data. Presented results rely on oracle information which will be replaced by a permutation solver in our future work.

The second contribution was a modification of the GSS system to enable the use of a diarization output as alternative to the human annotations for the second track of CHiME 6. Furthermore, we tested multiple acoustic models and experimented with model combination and an RNN language model. Our best system achieved the fourth best WER for the CHiME-6 Track 2 in the ranking for both constrained and unconstrained language models.

Table 3: *Submitted results. DER and Jaccard Error Rate (JER) are calculated against the provided alignments based targets.*

	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

9. Acknowledgements

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

10. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[2] 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 *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 11–15.

[3] N. Kanda, R. Ikeshita, S. Horiguchi, Y. Fujita, K. Nagamatsu, X. Wang, V. Manohar, N. E. Yalta 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 *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 6–10. [Online]. Available: http://dx.doi.org/10.21437/CHiME.2018-2

[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] 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.

[7] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[8] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, A. Kozlov, T. Pekhovsky, and Y. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," 2020.

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech 2017*, 2017, pp. 999–1003. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-620

[10] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[11] H. Daumé, A. Kumar, and A. Saha, "Frustratingly easy semi-supervised domain adaptation," in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, ser. DANLP 2010. USA: Association for Computational Linguistics, 2010, p. 53–59.

[12] 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.

[13] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 296–303.

[14] 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 the second dihard speech diarization challenge," 2020.

[15] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[24] 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.

[25] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[27] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[28] 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.

[29] C. Zorilă, M. Li, D. Hayakawa, M. Liu, N. Ding, and R. Doddipatla, "Toshiba's speech recognition system for the CHiME 2020 Challenge," in *Proc. of CHiME-6 Workshop*, 2020.

[30] 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.

[31] 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.