FRAME-WISE AND OVERLAP-ROBUST SPEAKER EMBEDDINGS FOR MEETING DIARIZATION

Tobias Cord-Landwehr[⋆], *Christoph Boeddeker*[⋆], *Cătălin Zorilă*[†], *Rama Doddipatla*[†], *Reinhold Haeb-Umbach*[⋆]

[⋆] Paderborn University, Department of Communications Engineering, Paderborn, Germany
[†] Toshiba Cambridge Research Laboratories, Cambridge, United Kingdom

ABSTRACT

Using a Teacher-Student training approach we developed a speaker embedding extraction system that outputs embeddings at frame rate. Given this high temporal resolution and the fact that the student produces sensible speaker embeddings even for segments with speech overlap, the frame-wise embeddings serve as an appropriate representation of the input speech signal for an end-to-end neural meeting diarization (EEND) system. We show in experiments that this representation helps mitigate a well-known problem of EEND systems: when increasing the number of speakers the diarization performance drop is significantly reduced. We also introduce block-wise processing to be able to diarize arbitrarily long meetings.

***Index Terms*—** speaker embeddings, diarization, EEND, d-vectors, teacher-student training, speaker verification

1. INTRODUCTION

Speaker embedding extraction aims at computing a representation from a speech signal that keeps the information about the speaker identity, while being insensitive to other factors, such as the content. Next to its obvious application in speaker recognition or verification tasks, there are other applications that benefit from such a voice fingerprint, such as Automatic Speech Recognition (ASR) [1], source extraction from a speech mixture [2] and speaker diarization [3, 4].

Early approaches to speaker embedding computation (e.g., the i-vectors [5]) employed statistical methods, however, more recently, the neural network-based techniques are prevalent. In the latter case, the speaker embedding is taken as the representation at a bottleneck layer right before the last, i.e., the classification layer of the network. X-vectors [6] are a popular neural network based representation that employs a Time Delay Neural Network (TDNN) with statistics pooling for computation. Another widely used speaker embeddings are the d-vectors that use ResNet-based convolutions for computation [7], achieving an impressive speaker verification performance on VoxCeleb [8].

Speaker embeddings have also been used for speaker diarization. For example, the baseline system of the third DiHARD challenge [3] employed an x-vector extractor, followed by similarity scoring and clustering. However, there are at least two issues when used in speaker diarization tasks. Firstly, the embeddings are computed on a relatively large signal window of typically 2 s to 4 s. This contradicts the benefit to have a high temporal resolution of the diarization information of a few frames. Secondly, embeddings computed from regions of overlapped speech, i.e., regions where more than one speaker is active, are typically uninformative. Ideally, the computed embedding vector for a mixture of two speakers should also be a superposition of the embeddings of the participating speakers. However, the embedding extraction is far from being linear, and the participating speakers cannot easily be inferred from the speech mixture. This is a well-known issue that hurts clustering-based approaches to speaker diarization.

In contrast, the End-to-End Neural Diarization (EEND) [9] method can handle speech overlap regions well by formulating diarization as a multi-label classification problem. However, EEND systems suffer from the fact that they solve this problem rather "locally", meaning that they can discern only a small number of speakers and they struggle to re-identify the speakers at a later point in time. To alleviate these issues, methods like EEND-EDA compute speaker-wise attractors to maintain a speaker state [10, 11], or employ ASR-generated features to achieve a similar goal [12].

Rather than fixing the problem posthoc by modifying the architecture of neural diarization, in this contribution we directly address the above two deficiencies of state-of-the-art speaker embedding extractors. This will in turn improve the speaker re-identification when using the embeddings as input to a neural diarization system.

Our proposal is to decouple the learning of the speaker embedding space from the projection of a speech signal into that space, using a Teacher-Student strategy. The teacher is a standard ResNet d-vector system, and the student is trained to output high temporal resolution (i.e. frame-level) vectors that replicate the teacher's embeddings. Therefore, the student projects a segment of input speech to the embedding vector representing the active speaker. While a pooling over several seconds is crucial for the teacher to learn a robust speaker embedding space, the student can work at a much higher temporal resolution. This leads to some performance drop in a large scale speaker verification task, but works sufficiently well for small numbers of speakers, as is typical for meeting diarization scenarios. Additionally, the frame-level resolution has the benefit that simultaneously active speakers are better recognizable from the embeddings computed from a mixture, because many frames are dominated by either of the two speakers. The pooling of the teacher, on the other hand, would average out the speaker-specific information in the mixture. Using the student embedding extractor as a front-end to a EEND system, we show that diarization performance is significantly improved on meeting data compared to the baseline EEND system.

In the next section we describe the architectures of the teacher and the student network as well as the Teacher-Student training. In Section 3 we motivate the usage of the student-embeddings as input features for neural speaker diarization. Then, the quality of the frame-wise student embeddings, both for single-speaker data and two-speaker mixtures, as well as their applicability for diarization, is evaluated in Section 4.

2. TEACHER-STUDENT EMBEDDINGS

The extraction of the speaker characteristics from an input speech signal using a neural network is typically achieved by introducing an information bottleneck right before the classification layer of the network. The network is trained to minimize a classification loss between the estimated and the true speaker label. Since the number of speakers seen during training is significantly larger than the dimensionality of the latent embedding space, the representation at the bottleneck layer is forced to represent the speaker characteristics and not only the speaker labels. In this way, speaker embedding extractors generalize well to unseen speakers [6].

Typically, a single embedding is computed from an utterance or from a few seconds of speech. This is done by Time Average Pooling (TAP) of the activations of the bottleneck layer. In fact, it has been shown that this aggregation over several seconds is necessary to achieve a good classification performance [13].

When employing these systems for diarization, the utterance is split into overlapping windows of 2 to 4 seconds length, from which one embedding vector per window is computed. However, the TAP renders the speaker embeddings unreliable if the environment changes over the duration of a window [14]. In this case, the aggregation of the activations does not deliver a meaningful speaker representation, as their properties change midway. This is a well-known issue of classical clustering-based diarization approaches [14, 15].

In order to obtain a speaker embedding extractor that does not require TAP and still computes reliable embeddings, we employ a Teacher-Student approach. The teacher is an ordinary speaker embedding extractor, and the student is the desired extractor that can produce reliable frame-wise embeddings, as is described next.

2.1. Teacher speaker embeddings

In this work, the teacher network is a typical ResNet-based d-vector extractor [16], that is widely used for speaker verification [17, 18]. Here, first logarithmic mel filterbank features $x(t, f)$ are extracted from an observation $x(\ell)$. These features are then encoded into frame-wise latent embeddings, which are then aggregated with a TAP layer to obtain a single, E-dimensional d-vector $d(e)$ representing the speaker characteristics of the input speech. During training, an additional fully connected classification layer is used to predict the speaker label. By using a softmax-based classification loss that normalizes the embeddings to unit length, the d-vectors lie on a E-dimensional hypersphere once the model is fully converged. However, this property only holds true for the d-vectors $d(e)$, not the frame-wise embeddings before the TAP. Other works [19] showed that these frame-wise embeddings heavily fluctuate in terms of power and can even be used to infer a Voice Activity Detection (VAD) from, so that no further assumptions about the structure of this data can be made.

2.2. Teacher-Student training

For the Teacher-Student training, a second network, the student, is trained while using the d-vectors $d(e)$ of the teacher as training targets as depicted in Figure 1. The weights of the teacher itself are kept fixed during this step. The architecture of the student network is largely identical to the teacher. Here, also a ResNet34 extracts frame-wise speaker embeddings from the logarithmic mel filterbank features $x(t, f)$ of the observation.

However, the student exhibits two key differences to the teacher. First, instead of performing a global TAP, a local TAP over 11 frames with a frame advance of single frame to smooth the output embeddings is applied. In this way, the student embeddings $\hat{d}(t, e)$ maintain a frame-wise resolution, while they still include the contextual frames during embedding extraction. Second, the student is not trained with a classification loss, but with a Mean Squared Error (MSE)-based similarity loss

$$\mathcal{L}_{\text{sim}} = \frac{1}{TE}\sum_{t=1}^{T}\sum_{e=1}^{E}\|d(e) - \hat{d}(t, e)\|^2 \quad (1)$$

between each frame-wise student embedding $\hat{d}(t, e)$ and the teacher d-vector $d(e)$. Through this loss, the student is not only encouraged to reproduce the d-vectors of the teacher, but to do so on a frame level. Therefore, every frame-wise student embedding ideally is projected onto the E-dimensional hypersphere that depicts the latent speaker space of the teacher. Those are easier to interpret, in particular in case of a speaker change or speech overlap, since they depict a higher time resolution.

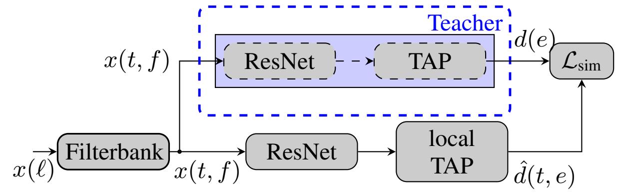


Fig. 1: Block diagram of the student training. The teacher d-vectors serve as targets for the frame-wise embeddings of the student.

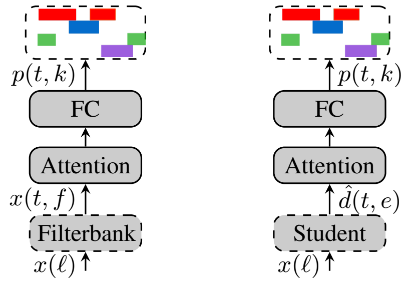
3. SPEAKER EMBEDDING-BASED EEND

To evaluate the previous hypothesis, we adapt the EEND approach [9]. In EEND systems, a neural network is trained to directly predict the activity of each speaker in a meeting recording from the observation. The EEND model consists of a feature extraction followed by self-attention layers and a classification layer to predict the frame-wise activity of each speaker. During inference, an additional min pooling (erosion) followed by a max pooling (dilation) is applied before thresholding the model output to obtain the estimated activities.

Here, we replace the feature extraction with the student embedding extractor as depicted in Figure 2 and train the EEND components while keeping the model weights of the student frozen. The resulting system is denoted “Student-EEND”. Since the speaker information is already encoded in the input features to the EEND model, the Student-EEND only needs to learn a multi label assignment of these input features to provide a good diarization, and the speaker information necessary to distinguish between speakers is directly available at the input.

3.1. Block-wise Student-EEND

For normal EEND models, the full length audio of the recording is required at once. In order to obtain a diarization system that can operate on arbitrarily long meetings, a block-wise processing is proposed in the following. Instead of processing the whole meeting at once, during inference the meeting is processed in overlapping blocks. Note that block processing results in a block permutation problem, i.e., the estimated activity $\bar{p}_b(t, k)$ after thresholding in each block b can be arbitrarily permuted. Block-wise processing has also been proposed in [20, 21]. In these works, the models are trained to output an additional state that can be used to solve the block permutation problem. Here, instead of inferring a speaker state from the



(a) Standard EEND [9]   (b) Student-EEND

Fig. 2: Comparison of the standard EEND and Student-EEND

EEND output, we directly use the student speaker embeddings to reorder the activity estimates and assign them to the correct speaker. Similarly to [20], this is done by first obtaining a block-wise d-vector by computing a weighted mean

$$d_b(k,e) = \frac{1}{T}\sum_{t=1}^{T} \bar{p}_b(t,k)\hat{d}_b(t,e) \tag{2}$$

for each active speaker k in the processed block b from the student embeddings $\hat{d}_b(t,e)$ with the estimated activities $\bar{p}_b(k,t)$. Then, inactive speakers are removed and the d-vectors of all blocks are clustered with an overlap-aware k-means clustering to find the block permutation. Next, the activities in the center parts of successive blocks are concatenated as in [22] to obtain the diarization estimate. The number of speakers active in a block tends to be smaller than the total number of speakers participating in a meeting. Therefore, the Student-EEND system can be trained for an expected maximal number of speakers in a block (in this work, maximum 4 speakers) and then applied to a meeting with arbitrarily many speakers. In this way, the Student-EEND can be used for a block-wise processing without any further modifications to the EEND model.

4. EVALUATION

4.1. Databases and training setup

The teacher and student models are trained using VoxCeleb data [8] augmented with noise from the MUSAN dataset [23] and RIRs from [24]. The teacher is trained with the additive angular margin (AAM) Softmax [25] loss and the student is trained with a frame-wise MSE loss to match the teacher's d-vectors.

Reverberant LibriSpeech meetings are simulated with the MMS-MSG software [26] which allows for the simulation of meetings with an arbitrary number of speakers and duration. These meetings contain $15\,\%$ to $20\,\%$ overlap and are used to evaluate the speaker embeddings under overlapping speech. Furthermore, the same data is employed to train and evaluate the EEND systems. $2\,\text{min}$ meetings are simulated and all the diarization models use $60\,\text{s}$ chunks for training. During evaluation, the Diarization Error Rate (DER), Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) are calculated according to [27].

4.2. Student embeddings for speaker verification

For the speaker verification performance, both the student and the teacher d-vector models are evaluated on the VoxCeleb speaker verification test sets. In this context, for better comparison with the teacher, the frame-wise embeddings $\hat{d}(t,e)$ of the student also are averaged over time to obtain a single student d-vector $\bar{d}(e)$. As can be seen in Table 1 , the student d-vectors consistently perform worse than the teacher d-vectors for this task. Finetuning the student model training by replacing the target d-vector with a different d-vector of the same speaker reduces the performance gap to some degree. The worse performance of the student is not surprising, as only few student models are able to outperform their teachers on their original tasks [28]. The student is, however, expected to perform reasonably well with a small number of speakers, as is the case for meetings.

Table 1: Speaker verification performance of the teacher and student models on the VoxCeleb test sets.

Model	VoxCeleb1		VoxCeleb1-H	
	EER[%]	minDCF	EER[%]	minDCF
Teacher	**1.06**	**0.08**	**2.71**	**0.08**
Student	2.34	0.19	5.33	0.29
+ finetuning	2.02	0.14	4.16	0.21

Table 2: Speaker identification accuracy of teacher and student d-vectors for single speaker and overlap regions

Model	Single speaker	Overlap
Teacher	0.98	0.35
Student	0.98	**0.42**

4.3. Embeddings on overlapped speech

For the purpose of conversational speech diarization, an important question is whether the embeddings are robust under overlapping speech. For this purpose, the teacher and student speaker embedding extractors are evaluated for 8-speaker LibriSpeech meetings. Both for the teacher and the student, short-time d-vectors with a duration of $2\,\text{s}$ and a hop size of $0.5\,\text{s}$ are extracted from these data. The lengths correspond to sizes typically chosen for speaker embedding based diarization systems. Then, prototype d-vectors for each speaker in the meeting are obtained from clean recordings, and they serve as a reference for the above segment d-vectors to identify the active speakers both in the single-speaker and overlapping regions. This is done by computing the cosine similarity between all prototypes and the segment d-vector.

Table 2 shows that student and teacher achieve the same accuracy for the identification of single speaker regions. As expected, in the overlap regions, however, the student's identification accuracy outperforms the teacher, indicating a higher robustness for overlapping speech.

Figure 3 explains why the student embeddings are able to better represent the active speakers for overlapping speech over the teacher embeddings. The figure displays the accumulated local cosine distance between successive frame-wise speaker embeddings. As shown, the distances for the student are much smaller than for the teacher. This indicates that the student's frame-wise embeddings smoothly change from one speaker to the next even during the overlap. We further observed that higher local distances coincide with speaker changes over the course of a meeting for the student.

This smooth transition can also be seen when visualizing the frame-wise embeddings with t-SNE. Figure 4 shows that, for the student, the embeddings computed from overlapped speech segments form trajectories connecting the embeddings computed from single-speaker regions of the speakers present in the mixture, whereas no

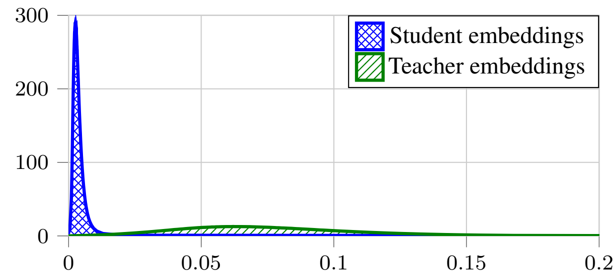
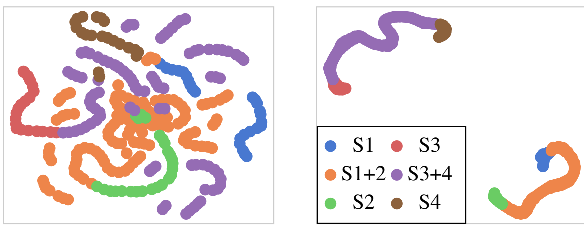


Fig. 3: Probability density of local cosine distance between consecutive frame-wise embeddings for reverberant 8-speaker meetings



(a) Teacher embeddings (b) Student embeddings

Fig. 4: t-SNE plot for the frame-wise embeddings of two segments consisting of partially overlapping speech.

discernible geometry can be seen for the teacher embeddings. Enforcing the student vectors to lie on the hypersphere defined by the teacher seems to help regularize the frame-wise speaker embeddings for overlapping speech, even if the student was not trained on speech mixtures. Note, that it cannot be concluded that these embeddings are linear combinations of the participating speaker embeddings, because t-SNE is a highly nonlinear mapping from a high-dimensional space to two dimensions.

4.4. Embeddings for diarization

The previous evaluations show that the frame-wise student embeddings are meaningful even under speech overlap. While simply assigning regions of overlapping speakers to the most similar speaker prototypes is still too unreliable, it should be possible to find a non-linear mapping of the frame-wise student embeddings which leads to an accurate diarization. Therefore, we use the Student-EEND model to this end and compare it to the standard EEND. The values for the activity threshold, erosion and dilation are finetuned separately for each system configuration. The training is performed using a permutation invariant binary cross entropy loss.

Table 3 shows that the student embeddings indeed can be used for diarization. For meetings with two speakers, both the standard EEND and Student-EEND perform similarly. However, when switching to meetings with four active speakers, the standard EEND steeply degrades, while the student embedding-based version still is able to provide a decent diarization result and clearly outperforms the EEND model. To verify that it is really the Teacher-Student training that is responsible for this effect, we also train an EEND model directly on the activations of the teacher d-vector system, i.e., on the embeddings before TAP, which we denote "Teacher-EEND". As shown, this configuration is unable to provide a meaningful diarization even for the two-speaker case. This again demonstrates that the student embeddings are easier to interpret on a frame-level and that they form a useful front-end for a neural diarization system.

Table 3: DER (Missed Hit/ False Alarm/ Confusion) on simulated LibriSpeech meetings with a duration of 2 min

Model	2 speakers	4 speakers
EEND	**6.1** (2.1 / 2.6 / 1.4)	33.2 (21.4 / 3.3 / 8.5)
Student-EEND	6.6 (2.4 / 1.9 / 2.3)	**13.4** (6.9/ 3.9 / 2.6)
Teacher-EEND	17.9 (6.0 / 8.5/ 3.4)	49.6 (0.4 / 32.9 / 16.3)

Table 4: DER for 10 min meetings with and without block-wise processing

Model	block-wise	#Speakers per Meeting 2	4	8
EEND	✗	**6.5**	38.8	-
Student-EEND	✗	7.6	**14.4**	-
Student-EEND (4spk)	✓	9.6	16.8	**30.9**

4.5. Block-wise Student-EEND

Next, we investigate the performance of the Student-EEND for a block-wise processing. Table 4 compares the block-wise approach with processing the full 10-mins meetings at once. The block-wise system is trained on 4-speaker meetings and evaluated on 30 s long blocks with a block advance of 10 s. The results for processing the full meeting have been obtained with systems that have been trained specifically for the given number of speakers.

It can be seen that the block processing leads to an increase in the DER, which is mainly due to the fact that the clustering could not resolve all block permutation errors. Furthermore, it can be observed that the least relative increase in DER is observed if the number of speakers in training matches those in test (i.e., 4). While the generalization to a smaller number of speakers is satisfactory, the deterioration for 8-speaker system to 31 % is more pronounced. This observed performance loss for more speakers is comparable to the degradation reported in [29]. Notably, training a 8-speaker EEND model was not possible due to the complexity associated with solving the permutation problem.

5. CONCLUSION

In this work, we proposed a Teacher-Student approach for the extraction of speaker embeddings. In contrast with the teacher, the student is able to produce sensible speaker embeddings at frame rate. Further, the student's speaker embeddings contain sufficient information for segmentation and speaker identification to be used as representation of the input speech for neural diarization systems, even during speech overlap. Experiments have shown that they help mitigate a well-known problem of EEND systems: when increasing the number of speakers in a meeting, the performance drop is much less pronounced. Furthermore, we presented a block-wise student-EEND system that can handle arbitrarily long meetings. Finally, it is worth mentioning that the student embeddings impose no constraints on the backend diarization system. Thus, more recent extensions of EEND-based systems (e.g. EEND-EDA [10]) can be easily combined with the Student-EEND, which serves as an outlook on future work.

6. ACKNOWLEDGEMENT

Computational resources were provided by the Paderborn Center for Parallel Computing. Christoph Boeddeker was funded by Deutsche Forschungsgemeinschaft (DFG), project no. 448568305.

7. REFERENCES

[1] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[2] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, 2019.

[3] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," in *Interspeech*, 2021.

[4] 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, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.

[5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, 2011.

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

[7] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an End-to-End neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

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

[9] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," in *Interspeech*, 2019.

[10] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for End-to-End neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, 2022.

[11] S. Horiguchi, S. Watanabe, P. García, Y. Xue, Y. Takashima, and Y. Kawaguchi, "Towards neural diarization for unlimited numbers of speakers using global and local attractors," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021.

[12] A. Khare, E. Han, Y. Yang, and A. Stolcke, "ASR-aware End-to-End neural diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[13] 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," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2020.

[14] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, 2022.

[15] D. Raj, L. Paola Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A method for combining overlap-aware diarization outputs," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[16] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net structures for speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[17] X. Qin, N. Li, Y. Lin, Y. Ding, C. Weng, D. Su, and M. Li, "The DKU-Tencent system for the VoxCeleb speaker recognition challenge 2022," *arXiv preprint arXiv:2210.05092*, 2022.

[18] S. Suh and S. Park, "The ReturnZero system for VoxCeleb speaker recognition challenge 2022," *arXiv preprint arXiv:2209.10147*, 2022.

[19] Y. Kwon, H. S. Heo, J. Huh, B.-J. Lee, and J. S. Chung, "Look who's not talking," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[20] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating End-to-End neural and clustering-based diarization: Getting the best of both worlds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[21] Y. Xue, S. Horiguchi, Y. Fujita, Y. Takashima, S. Watanabe, L. P. G. Perera, and K. Nagamatsu, "Online Streaming End-to-End Neural Diarization Handling Overlapping Speech and Flexible Numbers of Speakers," in *Interspeech*, 2021.

[22] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous Speech Separation: Dataset and Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[25] Y. Liu, L. He, and J. Liu, "Large Margin Softmax Loss for Speaker Verification," in *Interspeech*, 2019.

[26] T. Cord-Landwehr, T. von Neumann, C. Boeddeker, and R. Haeb-Umbach, "MMS-MSG: A multi-purpose multi-speaker mixture signal generator," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. 2022, IEEE.

[27] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 speaker recognition evaluation plan," 2021-07-12 04:07:00 2021.

[28] G. Hinton, O. Vinyals, J. Dean, et al., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[29] K. Kinoshita, M. Delcroix, and T. Iwata, "Tight integration of neural-and clustering-based diarization through deep unfolding of infinite gaussian mixture model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.