INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Teacher-Student approach for extracting informative speaker embeddings from speech mixtures

Tobias Cord-Landwehr[1], *Christoph Boeddeker*[1], *Cătălin Zorilă*[2], *Rama Doddipatla*[2], *Reinhold Haeb-Umbach*[1]

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

cord@nt.upb.de, boeddeker@nt.upb.de, catalin.zorila@toshiba.eu, rama.doddipatla@toshiba.eu, haeb@nt.upb.de

Abstract

We introduce a monaural neural speaker embeddings extractor that computes an embedding for each speaker present in a speech mixture. To allow for supervised training, a teacher-student approach is employed: the teacher computes the target embeddings from each speaker's utterance before the utterances are added to form the mixture, and the student embedding extractor is then tasked to reproduce those embeddings from the speech mixture at its input. The system much more reliably verifies the presence or absence of a given speaker in a mixture than a conventional speaker embedding extractor, and even exhibits comparable performance to a multi-channel approach that exploits spatial information for embedding extraction. Further, it is shown that a speaker embedding computed from a mixture can be used to check for the presence of that speaker in another mixture.

Index Terms: speaker embeddings, teacher-student, multi-speaker, speaker identification, monaural

1. Introduction

Speaker embeddings are meant to represent the characteristics of a speaker, while being insensitive to what has been spoken. The embeddings are computed from segments of speech in such a way that they exhibit low intra-speaker and large inter-speaker distance. The current state of the art are neural speaker embedding extractors. They map the input speech segment to the latent space of embedding vectors by means of a neural network, which can be trained using a contrastive [1] or classification-based loss [2–4]. Since the number of speakers seen during training is much larger than the dimension of the latent space, the model is forced to encode the speaker characteristics rather than memorizing the speaker labels in the speaker embeddings. In this way, speaker embedding extractors generalize well to speakers unseen during training [5]. Speaker embedding extractors are commonly employed either as auxiliary systems for speech enhancement [6] or automatic speech recognition [7], as part of a diarization system [8], or as standalone systems for the (re-)identification and recognition of speakers [9].

An assumption underlying most current systems is that each processed audio segment contains only a single speaker of interest. If this assumption is not fulfilled, it is well known that the quality of such speaker embeddings degrades, and that the resulting speaker embedding is at most able to represent one of the speakers in the mixture. Because of this, diarization systems that employ speaker embedding extractors typically discard regions that contain speech overlap [10], or they use an additional uncertainty state to account for the lower reliability of speaker information extracted from those regions of speech [11, 12]. Then, the speakers active in those regions often are inferred from the context. The extraction of reliable speaker embeddings thus rests on the availability of regions where the speaker of interest is active alone. Even state of the art diarization systems like the TS-VAD [13] have this dependency on long single-speaker regions for proper initialization [14].

In highly dynamic situations, e.g. in informal meetings or in situations where multiple separate conversations are held in parallel, the requirement that each speaker is at least once solely active, cannot be fulfilled easily. Here, at least some speakers will not appear alone, so that a typical speaker embedding extractor cannot be used to extract a representation of those speakers. One way to approach this issue is to employ a source separation module as a preprocessor to the standard speaker embedding extractor [15], which, however, can introduce additional artefacts into the signal. Alternatively, a multi-channel system is proposed in [16], which computes features for each sound direction of arrival to detect and identify speakers.

In this work, we neither require a source separation frontend nor multi-channel input. Instead, we aim to directly compute from a mixture of two speech signals their respective speaker embeddings. In a sense, this is similar to source separation since we extract two embeddings from the input mixture. However, there is a clear difference: unlike the speech signals, the speaker embeddings cannot be assumed to linearly superpose in a mixture. As a consequence, a specific setup is required for training. We employ the teacher-student approach introduced in our previous work [17]: First, a conventional speaker embedding extractor is trained on single-speaker utterances. Then, this model is employed as a teacher for the training of a student embedding extractor. The student's objective is to predict those embedding vectors from a segment of speech mixture at its input. One way to view this setup is that the teacher defines the latent space of speaker embeddings, while the student inherits this space and is only tasked to separate a mixture of two speakers into their respective representations in that space. Contrary to the Wavesplit source separation model [18], where a speaker representation is jointly learned with the separation task, the latent space of the speaker embeddings already is explicitly provided by the teacher. Therefore, similar to the problem of source separation, we view this system as a speaker embedding separation approach.

This work is structured as follows: We start with a description of both the teacher and student models in section 2. Then, section 3 describes the task of multi-speaker verification and how we trace it back to the classical speaker verification task. In section 4, the proposed model is then evaluated against a typical single-speaker embedding extractor on a multi-speaker verification task. Finally, we compare our system against a multi-speaker embedding extractor on realistic meeting recordings.

10.21437/Interspeech.2023-1379

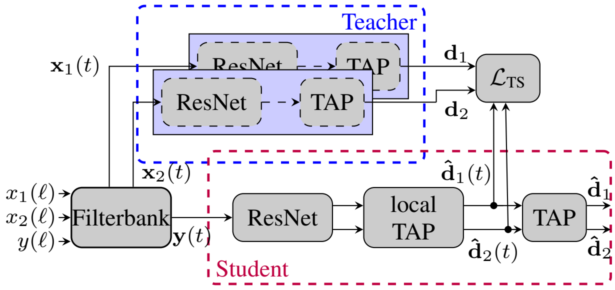


Figure 1: *Block diagram of the proposed teacher-student training for multi-speaker embedding extraction*

2. Teacher-student based embedding extraction

The proposed system for multi-speaker embedding extraction consists of a single-speaker teacher that is used to provide training targets for a multi-speaker student model. First, the teacher model is trained with single-speaker utterances. Then, the teacher model is fixed, and the student network is trained with speech mixtures at its input and the speaker embeddings of the teacher as training targets.

2.1. Single-speaker embedding extractor

Essentially, any speaker embedding extractor can be used as a teacher. In this work, a ResNet34-based d-vector system [4] is employed, which is a widely used architecture for speaker embedding extractors [19–21]. First, 80-dimensional log mel filterbank features $\mathbf{x}(t)$, $t \in \{1, \ldots, T\}$, with frame index t and frames per utterance T are extracted from an utterance $x(\ell)$ with time domain sample index ℓ. These features are passed through the ResNet to obtain frame-wise embeddings $\mathbf{d}(t)$. Then, the embeddings are aggregated by a Time Average Pooling (TAP) to obtain a single, E-dimensional d-vector $\mathbf{d}$ for the utterance. During training, an additional fully connected layer is used to predict the speaker label c from this d-vector. An Additive Angular Margin (AAM)-Softmax loss [22]

$$\cos \Theta_c = \frac{\mathbf{d}^{\mathrm{T}} \mathbf{w}_c}{|\mathbf{d}||\mathbf{w}_c|} \tag{1}$$

$$\mathcal{L}_{\mathrm{AAM}} = -\log \frac{e^{s(\cos \Theta_c - a)}}{e^{s(\cos \Theta_c - a)} + \sum_{c'=1, c' \neq c}^{C} e^{s \cos \Theta_{c'}}} \tag{2}$$

is used as a training criterion. Here, Θ_c denotes the angle between the d-vector $\mathbf{d}$ and the prototype embedding of the corresponding speaker which is defined by the c-th column $\mathbf{w}_c$ of the weight matrix of the fully connected layer. C represents the total number of speakers during training. This loss is a softmax cross-entropy loss, where the hyperparameters $s > 1$ and $a > 0$ modify the loss contributions such that the model builds more compact clusters after convergence, which improves the overall system performance [2]. After training, this single-speaker model is kept fixed and the d-vectors $\mathbf{d}$ are used as targets to train the student.

2.2. Multi-speaker embedding extraction

For the training of the multi-speaker student system, a speech mixture $y(\ell)$ consisting of K single speaker source signals $x_k(\ell)$ (here, $K = 2$) is used as input for the system. Then, the student is tasked to reproduce the d-vectors that are extracted by the teacher from the single speaker utterances $x_k(\ell)$.

The multi-speaker student's architecture is in large parts a mirror of the teacher. First, 80-dimensional log mel filterbank features $\mathbf{y}(t)$ are extracted from the speech mixture. Then, they are passed through a ResNet34 with a K times larger output dimension and are rearranged to obtain frame-wise speaker embeddings $\hat{\mathbf{d}}_k(t)$ for each speaker. Additionally, these embeddings are then averaged by a local TAP layer with a size of 11 and a stride of 1 to encode context information into each embedding. Then, a Mean Squared Error (MSE)-based similarity loss

$$\mathcal{L}_{\mathrm{TS}} = \frac{1}{KT} \sum_t \min_{\pi \in \mathcal{P}} \sum_k \|\mathbf{d}_k - \hat{\mathbf{d}}_{\pi_k}(t)\|^2 \tag{3}$$

is computed between each frame-wise student embedding $\hat{\mathbf{d}}_k(t)$ and the teacher embeddings $\mathbf{d}_k$ obtained from the single-speaker observations.

To account for the arbitrary output order of the student embeddings, a Permutation Invariant Training (PIT) loss [23, 24] is used, which assigns the best permutation π between target and estimated d-vectors for each frame (tPIT). Alternatively, the permutation can be kept constant for the complete utterance by moving the $\min$ operation in (3) in front of the sum over the frames t, which is known as uPIT. During inference, an additional TAP layer aggregates the frame-wise student embeddings into a single embedding $\hat{\mathbf{d}}_k$ per speaker.

With this teacher-student training, instead of having to learn a latent space as well as to separate both active speakers from each other, the student is trained to directly reproduce the single speaker embeddings of the teacher, because the latent speaker embedding space is already defined by the teacher. This simplifies the problem from one of learning a descriptive speaker space while separating speakers from each other to a task of speaker embedding separation into an already known, latent space. An illustration of the complete teacher-student model is depicted in fig. 1.

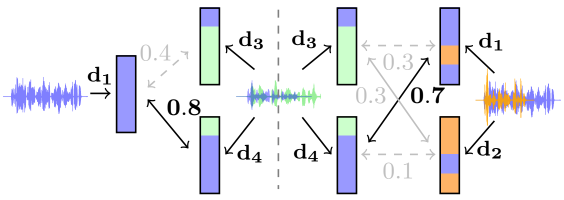
3. Speaker verification for overlapping speech

Typically, the task of speaker verification [25] consists of computing the similarity of two single-speaker utterances and deciding whether they belong to the same speaker. After obtaining these similarity scores for a complete trial set, they are aggregated into a single list and compared against the target labels. Then, the Equal Error Rate (EER) and Detection Cost Function (DCF) are calculated based on the scores and labels to evaluate the speaker verification performance [25].

However, this task is no longer clearly defined if one or both of the observations in such a trial pair contain more than one speaker. Similar to [16] we extend the classical verification task "single speaker vs. single speaker" (*s vs. s*) by two additional scenarios to also consider speech mixtures:

- single speaker vs. mixture (*s vs. m*)
- mixture vs. mixture (*m vs. m*).

For multi-speaker embedding extraction, the *s vs. m* setting is the most relevant scenario. In most settings, e.g. a meeting or conversation, each speaker is the sole active speaker at least at some point in time. Therefore, it can be assumed that these single-speaker regions can be used to extract embeddings that are then taken as reference to verify whether this speaker is active in regions containing overlapping speech [26]. On the other hand, the *m vs. m* scenario is relevant for scenarios, where even



a) s vs. m evaluation *b) m vs. m evaluation*

Figure 2: *Multi-speaker verification process. For each active speech component, an embedding is extracted, and all pairwise similarities are computed. Only the maximal score is retained for "any spk" evaluations. For the "per spk" case, the remaining embedding pair also is evaluated in the m vs m scenario.*

this constraint cannot be fulfilled, which may happen, e.g., if multiple simultaneous discussions are ongoing at the same time.

First, we evaluate both scenarios with the goal to determine whether *any speakers* are matching between both examples in a trial pair. Here, from each observation embeddings are extracted as depicted in fig. 2, and all pairwise similarity scores between the first half and the second half of a trial pair are calculated, i.e., 2 and 4 in the *s vs. m* and the *m vs. m* setting, respectively, for $K = 2$. Then, only the maximal similarity score is retained for evaluation. The calculation of EER and DCF then is done exactly as for the single-speaker case. Since only a single similarity score is evaluated, typical single-speaker embedding extractors can also be evaluated in this setup, even though they are not designed for it.

For the *m vs. m* scenario, in addition to this "*any spk*" evaluation, the mixtures are compared *per speaker* to measure how well both speakers of a mixture can be represented. Here, the number of positive target labels is determined through the number of identical speakers between both mixtures. These labels are then assigned to the embedding pairs in a greedy fashion beginning by the highest similarity and excluding assigned pairs for following labels. Therefore, for each trial pair, K scores and target labels are obtained, as opposed to a single one in the "*any spk*" setting. They are then again aggregated and evaluated as in the *s vs. s* evaluation to compute both the EER and DCF.

4. Evaluation

4.1. Training setup

For training of both the teacher and the student model, 4 s long segments of the VoxCeleb corpus [27] are used. For filterbank feature extraction, a window size of 20 ms and a frame advance of 8 ms are chosen. The teacher is trained on single-speaker utterances augmented with noise from the MUSAN corpus [28] and room impulse responses simulated according to [29]. The student is trained on speech mixtures simulated with MMS-MSG [30] that consist of two speakers mixed with a power ratio between -5 dB to 5 dB. During training, these utterances are cut to the *min* scenario, i.e. the longer utterance in the mixture is cut to match the length of the shorter. Again, the same data augmentation as for the teacher is used. Additionally, after 10 epochs of training, the target embeddings provided by the teacher are replaced with embeddings computed from different utterances of the same speaker to make the model more robust against remaining speaker-unrelated content.

Table 1: *Speaker verification performance of the teacher model on the VoxCeleb1-O trial set (s vs. s)*

Model	E	EER [%]	DCF
ECAPA-TDNN [3]	1024	0.87	0.11
ECAPA-TDNN [3]	512	1.01	0.13
ResNet34	256	1.06	0.16

Table 2: *Multi-speaker verification performance for the student model and a multi-speaker embedding extractor trained with a classification loss on the VoxCeleb1-O m vs. m trials*

Loss	Permutation	EER [%]	DCF
$\mathcal{L}_{\mathrm{AAM}}$	uPIT	29.4	1.0
$\mathcal{L}_{\mathrm{AAM}}$	tPIT	29.5	1.0
$\mathcal{L}_{\mathrm{TS}}$	uPIT	19.9	0.88
$\mathcal{L}_{\mathrm{TS}}$	tPIT	**14.1**	**0.74**

4.2. Evaluation sets

For evaluation, the VoxCeleb1-O trial set [27] is extended by additional examples so that the *s vs. m* and *m vs. m* scenarios can be evaluated[1]. Here, still at most one speaker is identical in the two halves of a trial pair as to keep them as close as possible to the original VoxCeleb1-O trial set and not introduce the number of identical speakers as an additional design parameter in the trial sampling.

Additionally, the proposed model is evaluated on the SSLR database [31]. This database was originally designed for source localization and consists of re-recordings of AMI meetings [32]. It was already used in [16] for the evaluation of multi-channel multi-speaker embeddings with the same scenarios as described in section 3. All models are evaluated w.r.t. EER and DCF. For DCF calculation, a prior probability of 0.01 [25] is chosen for the single speaker evaluation, and of 0.05 for the multi-speaker verification.

4.3. Performance of the teacher model

The proposed teacher-student approach requires high-quality teacher embeddings that are used as target. Table 1 shows the teacher model's performance on the VoxCeleb1-O trial set compared to the popular ECAPA-TDNN [3]. Here, it can be seen that the ResNet is able to achieve a comparable performance. To keep a frame-wise resolution for loss computation, the ResNet architecture is the better choice for the student. Although any speaker embedding extractor can be used as teacher, we also chose a ResNet-based teacher to have matching configurations for teacher and student.

4.4. Necessity of teacher-student training

First, we compare our proposed teacher-student training with a multi-speaker model that is trained from scratch using the same classification loss as the teacher, albeit with a PIT objective. This second training approach is reminiscent of the way neural speech separation models are trained. Both models, the student of the T/S-approach and the extractor trained from scratch, use the same architecture described in section 2 and only differ in the loss used for training. Table 2 shows that training the speaker embedding extractor with a classification-based loss does not lead to a meaningful representation of the speaker embeddings independent of whether the permutation is solved per

[1]Trials are available at https://zenodo.org/record/7683872

Table 3: *Multi-speaker verification performance of the proposed model compared to a classical speaker embedding extractor on VoxCeleb mixtures. Model 1 and 2 indicate the model used on the first and second half of a trial pair.*

Model 1	Model 2	Scenario	any spk		per spk	
			EER[%]	DCF	EER[%]	DCF
Teacher	Teacher	s vs. m	18.2	0.57	-	-
Teacher	Student	s vs. m	**9.1**	**0.46**	-	-
Teacher	Teacher	m vs. m	47.6	1.0	-	-
Student	Student	m vs. m	**15.3**	**0.74**	**14.1**	**0.74**

frame or per utterance. On the contrary, the teacher-student approach results in an EER of 14.1 %. Using a training loss not on a frame-, but on an utterance-level did not work for either loss function. This shows that the extraction of two embeddings from a speech mixture is considerably more intricate than the extraction of the speakers' speech signals from the mixture. This is probably because while the mixture is a linear superposition of the speech signals, it is not so for the embeddings. Jointly learning a latent space and then projecting an utterance into this space leaves too many degrees of freedom for speech mixtures. However, by providing this latent space through a pretrained speaker embedding extractor, the (student) model extracts significantly better speaker embeddings that can be used to distinguish between the speakers.

4.5. Speaker embedding extraction for speech mixtures

Next, we compare the advantage of using a distinct multi-speaker embedding extractor over using only a single-speaker embedding extractor even in the presence of overlapping speech. Here, both the teacher and proposed model are evaluated on the extended VoxCeleb1-O trial sets. For the proposed model, the teacher is used to extract embeddings from single-speaker utterances, and the student for the embedding extraction from speech mixtures. Table 3 depicts the advantage of the combination of teacher and student over only using a single-speaker extractor. As expected, the proposed model outperforms the teacher in all scenarios. For the *s vs. m* trials, the teacher is still able to re-identify speakers in a mixture to some degree, but its performance drops sharply when switching to the *m vs. m* scenario where single-speaker regions are no longer available. On the contrary, the student is able to achieve an EER of 15.3 %, which is, however, significantly higher compared to the *s vs. m* scenario. In the "per spk" evaluation, the error rate decreases slightly, indicating that both extracted speaker embeddings can be used to identify the respective speakers.

To further quantify where the improvement of the student comes from and how disjoint the student embeddings are from each other, table 4 evaluates how well the embeddings extracted from the mixture represent each active speaker. This is done by calculating the cosine similarity between the speaker embeddings extracted from the mixture ($\hat{\mathbf{d}}_k$) and the respective teacher embeddings extracted from the clean, single-speaker signals ($\mathbf{d}_k$). This investigation shows that the teacher is able to accurately represent a single, the dominant, speaker in a mixture ($\mathbf{d}_y$), whereas the similarity to the other speaker stays low. Compared to that, the student network shows a slightly lower, albeit still high average similarity to the dominant speaker, and an increased similarity to the second speaker. Therefore, the student is able to extract both speakers, not only the dominant one as the teacher does.

Table 4: *Mean cosine similarity between different embeddings in a mixture and standard deviation over the trial examples (i.e. not model standard deviation). Similarities are ordered such that the higher score always is assigned to the first speaker.*

	Student (on y)		Teacher		
Teacher	$\hat{\mathbf{d}}_1$	$\hat{\mathbf{d}}_2$	$\mathbf{d}_1$	$\mathbf{d}_2$	$\mathbf{d}_y$
$\mathbf{d}_1$	.67 ± .09	.19 ± .22	1	-.09 ± .12	**.78 ± .14**
$\mathbf{d}_2$	.13 ± .19	**.40 ± .18**	-.09 ± .12	1	.22 ± .18

Table 5: *EER (per spk) on the different SSLR trial sets for varying maximal segment lengths. * marks multi-channel models.*

Model	s vs. m			m vs. m		
	10 s	5 s	2 s	10 s	5 s	2 s
MVDR+x-vector * [16]	10.8	12.7	20.4	12.2	14.3	21.7
Hybrid Model* [16]	**6.4**	10.2	18.4	**6.3**	**10.8**	19.3
Proposed	8.4	**9.1**	**14.2**	10.9	11.7	**18.2**

4.6. Evaluation on AMI re-recordings

Finally, the teacher-student model is evaluated on the AMI re-recordings from the SSLR dataset [31]. Here, the trial sets for multi-speaker verification consist of all possible pairwise segment combinations per room. Table 5 shows that our proposed single-channel system consistently outperforms the multi-channel baseline from [16], which is a multi-channel system consisting of an MVDR beamformer and an x-vector embedding extractor, in the *s vs. m* scenario, and in some cases also the hybrid model proposed in [16]. This effect becomes more pronounced for shorter segment lengths and can also be seen in the *m vs. m* scenario. For longer segments, the direction of arrival information used in [16] proves to be effective. Also noteworthy is that the SSLR dataset only contains 286 two-speaker segments in total, so more similar evaluation data may be necessary to solidify the results for the *m vs. m* scenario. Nevertheless, the results so far show that the proposed model achieves very good performance in the *s vs. m* scenario on realistic monoaural speech data without any finetuning.

5. Conclusions

In this work, we proposed a system for speaker embeddings extraction from speech mixtures. Using an embedding space defined by the teacher, a student embeddings extractor is learnt to cast a mixture input to embeddings in that space, representing the speakers present in the mixture. Thus, speakers that have previously been active in some speech segments as the sole speaker, can be tested for their presence or absence in a mixture. The (re-)identification even works, however less reliably, if the speaker has never been active alone before. On re-recordings of the AMI dataset the proposed approach is able to outperform a multi-channel approach without additional finetuning, at least for short segments. As future work, we are planning to use the proposed multi-speaker embedding extractor to derive a speaker diarization system.

6. Acknowledgements

Computational resources were provided by BMBF/NHR/PC2. Christoph Boeddeker was funded by DFG (German Research Foundation) under project no. 448568305.

7. References

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

[2] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech 2019*. IEEE, 2019, p. 2873–2877.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*. IEEE, 2020, pp. 3830–3834.

[4] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 301–307.

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

[6] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6099–6103.

[7] 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.

[8] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 897–904.

[9] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge," *ArXiv*, 2022.

[10] H. Bredin and A. Laurent, "End-To-End Speaker Segmentation for Overlap-Aware Resegmentation," in *Proc. Interspeech 2021*, 2021, pp. 3111–3115.

[11] D. Raj, Z. Huang, and S. Khudanpur, "Multi-class spectral clustering with overlaps for speaker diarization," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 582–589.

[12] 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," *Comput. Speech Lang.*, vol. 71, Jan 2022. [Online]. Available: https://doi.org/10.1016/j.csl.2021.101254

[13] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. Interspeech 2020*. IEEE, 2020, pp. 274–278. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1602

[14] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, "The DKU-DukeECE-Lenovo System for the Diarization Task of the 2021 VoxCeleb Speaker Recognition Challenge," *ArXiv*, vol. abs/2109.02002, 2021.

[15] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, S. Chen, Y. Zhao, G. Liu, Y. Wu, J. Wu *et al.*, "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5824–5828.

[16] W. He, P. Motlicek, and J.-M. Odobez, "Multi-Task Neural Network for Robust Multiple Speaker Embedding Extraction," in *Proc. Interspeech 2021*. IEEE, 2021, pp. 506–510.

[17] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Frame-wise and overlap-robust speaker embeddings for meeting diarization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.

[19] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "SJTU-AISPEECH System for VoxCeleb Speaker Recognition Challenge 2022," *arXiv preprint arXiv:2209.09076*, 2022.

[20] X. Qin, N. Li, C. Weng, D. Su, and M. Li, "Simple Attention Module Based Speaker Verification with Iterative Noisy Label Detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6722–6726.

[21] J. Thienpondt, B. Desplanques, and K. Demuynck, "Tackling the Score Shift in Cross-Lingual Speaker Verification by Exploiting Language Information," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7187–7191.

[22] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

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

[24] M. Kolbaek, D. Yu, Z. Tan, and J. H. 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, pp. 1901–1913, 2017.

[25] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 Speaker Recognition Evaluation Plan," 2021-07-12 04:07:00 2021. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932697

[26] A. Aloradi, W. Mack, M. Elminshawi, and E. A. P. Habets, "Speaker Verification in Multi-Speaker Environments Using Temporal Feature Fusion," in *30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 354–358.

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

[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[30] 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)*. IEEE, 2022.

[31] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[32] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The AMI meeting corpus," 2005.