



Diarization-Guided Multi-Speaker Embeddings

Joonas Kalda¹, Clément Pagés², Tanel Alumäe¹, Hervé Bredin²

¹Tallinn University of Technology, Estonia

²IRIT, Université de Toulouse, CNRS, Toulouse, France

joonas.kalda@taltech.ee

Abstract

Reliable speaker embeddings are critical for multi-speaker speech processing tasks. Traditionally models are trained on single-speaker utterances and suffer from domain mismatch when applied in multi-speaker contexts. Recently proposed guided speaker embeddings (GSE) were shown to improve this by training on synthetic multi-speaker mixtures guided by oracle speaker activity labels. Additionally modeling all speakers present in a chunk is desirable but the performance of such methods has been sub-par up to now. We build on GSE by modeling multiple speakers together and using diarization features for guiding. We also propose a new validation metric for embeddings in multi-speaker context and demonstrate its effectiveness. Results on multiple speaker diarization datasets demonstrate that we improve on speed and performance while reducing the embedding model size.

Index Terms: speaker diarization, speaker verification

1. Introduction

High-quality speaker embeddings are essential for multi-speaker speech processing tasks. In speaker diarization, EEND-vector clustering (EEND-VC) systems rely on speaker embeddings derived from local segmentation outputs to cluster and stitch these local segments [1]. Similarly, in multi-speaker automatic speech recognition (ASR), transducer-based systems produce segment-wise transcriptions with timestamp estimates, which are subsequently attributed to individual speakers using speaker embeddings [2]. For voice conversion of long-form audio, it is desirable that speaker embeddings are modelled consistently [3]. In all the above use cases, it would be beneficial for all the speakers to be modeled concurrently.

Previous studies have explored joint training of ASR and segmentation models with multi-speaker embeddings [2, 4, 5], but these approaches have underperformed compared to standalone embedding systems [6]. This discrepancy is likely due to the difference in data quality: speaker verification datasets, which are more easily annotated, tend to be larger and more diverse than those used for ASR and speaker diarization [7]. These datasets contain only single-speaker utterances, leading to a domain mismatch when applied in multi-speaker scenarios. To address this, guided speaker embeddings (GSE) were recently introduced for multi-speaker environments [8]. GSE is trained on synthetic multi-speaker mixtures derived from a speaker verification dataset, with oracle activity labels guiding the process. Activity labels for both target and interference speakers are used as additional inputs to the embedding encoder and for masking attention scores. However, these systems still produce embeddings for only one speaker at a time.

This work proposes extending GSE by modeling all speak-

ers present in a chunk at once. Additionally, since the practical deployment of a GSE system relies on a speaker segmentation model, we also propose using its output as a guide for training instead of oracle labels. Features from a segmentation model can offer more detailed guidance. For example, areas of high confidence indicate it is easier to extract speaker-specific information there.

The main contributions of this work are as follows:

- Proposing a diarization-guided training method for multi-speaker embedding systems.
- Introducing a modified attention module to allow for multi-speaker modeling in existing speaker embedding models.
- Proposing a new validation metric optimized for speaker embeddings in a multi-speaker context.
- Providing a thorough evaluation of the multi-speaker embeddings on multiple speaker diarization and verification datasets.
- Open-sourcing the code for the above¹.

2. Method

Figure 1 illustrates our joint architecture, which combines a local speaker segmentation model with a speaker embedding model using a shared feature extractor. We opt for an SSL-based features extractor, namely WavLM, since it demonstrates good performance in both speaker diarization and speaker verification tasks [9]. It is also the choice for state-of-the-art for speaker diarization as of writing [10, 11]. We use the same LSTM-based segmentation probing head as in [11]. For the embedding module, we use an ECAPA-TDNN, which has been shown to perform better than smaller probing heads [12, 13].

Given the frame-level features extracted from an audio chunk $\mathbf{x} \in \mathbb{R}^{T \times F}$ and assuming a maximum of $K_{\max}$ speakers, the segmentation module extracts powerset features $\mathbf{p} \in \mathbb{R}^{T \times K_{\text{ps}}}$, where K_{ps} is the number of powerset classes. These are binarized and converted into a multi-label format $\mathbf{a} \in \{0, 1\}^{T \times K_{\max}}$ [14]. The powerset features are concatenated with $\mathbf{x}$ to form combined features of dimension $F + K_{\text{ps}}$, which are fed into the embedding encoder. Only the input channel dimension of the encoder is modified in our approach.

The encoder output $\mathbf{h} \in \mathbb{R}^{T \times D}$ is reshaped to introduce a speaker dimension, resulting in $\mathbf{h}' \in \mathbb{R}^{K_{\max} \times T \times (D/K_{\max})}$. The attention module remains unchanged from the original ECAPA-TDNN, except that all the channel dimensions are scaled down by a factor of $K_{\max}$, except for the bottleneck attention dimension, which is kept at 128. The batch size after the encoder is effectively increased $K_{\max}$ times, with the speakers being processed in parallel.

¹<https://github.com/joonaskalda/multi-speaker-embeddings>

Similarly to GSE, we apply silent masking for each predicted speaker but use binarized predicted speaker activations instead of oracle labels. The embedding dimension for the predicted multi-speaker embeddings $\{\mathbf{e}_1, \dots, \mathbf{e}_{K_{\max}}\} \in \mathbb{R}^{192}$ is kept unchanged from the original ECAPA-TDNN.

Note that this approach would require slight modifications if the encoder output channel dimension D is not divisible by $K_{\max}$ by e.g. adding an adaptation layer. In the above we also assumed, for simplicity, that the embedding encoder leaves the temporal resolution unchanged, as is the case for ECAPA-TDNN, used in our experiments. If that is not the case, the speaker activation masks should be interpolated to match the embedding output temporal resolution.

To train the multi-speaker embedding model we use synthetic VoxCeleb mixtures as in [8]. We use the standard ArcFace loss [15] but only compute it for an embedding if the segmentation model correctly predicts the corresponding speaker's activation for at least one second.

In our experiments, we use a segmentation probing head trained using powerset loss [14], but this is not a requirement. The only new components that need to be trained in our method are in the speaker embedding branch. For the segmentation branch and feature extractor, any off-the-shelf model can be utilized, and no specific adaptation is needed. Our proposed training method generalizes naturally to any local segmentation and speaker embedding architecture, with no requirement for a shared feature extractor.

To summarize, our method builds on top of GSE by

- Changing the attention module to extract multi-speaker embeddings instead of single-speaker embeddings.
- Utilizing detailed information from the segmentation module.

2.1. Validation metric

The standard validation metric for speaker embedding models is the equal error rate (EER), computed on single-speaker utterance trials. However, this does not reflect performance in multi-speaker scenarios. We argue that speaker diarization performance is a more appropriate metric, as the clustering stage directly depends on embedding quality. We therefore propose using diarization performance for both evaluation and validation. Validation is challenging; normally, it would require hyperparameter optimization after each epoch, which is a resource-intensive process.

To address this, we propose a simplified pipeline for validation (Figure 2). In each validation batch, all audio chunks are sampled from the same file using a sliding window, ensuring that the first chunk starts at the beginning of the audio and the last one stops at the file's end. The step size S between consecutive chunks is chosen so that all chunks are evenly spaced i.e. $S = \frac{D-T}{B-1}$, where D is the file duration, B is the batch size, and T is the chunk length. Chunks overlap if and only if $BT > D$.

The batch is then fed to the model, which returns segmentation predictions and speaker embeddings for each chunk. These embeddings are then clustered using the K-means algorithm, where the number of clusters is fixed based on the oracle number of speakers in the file used to create the batch. For efficiency, we assume the number of speakers is known, which eliminates the need for tuning clustering parameters. Finally, segments are assigned to speakers as in a standard diarization pipeline, and a diarization error rate (DER) over the validation files is calculated based on the pipeline output and the corresponding cropped reference.

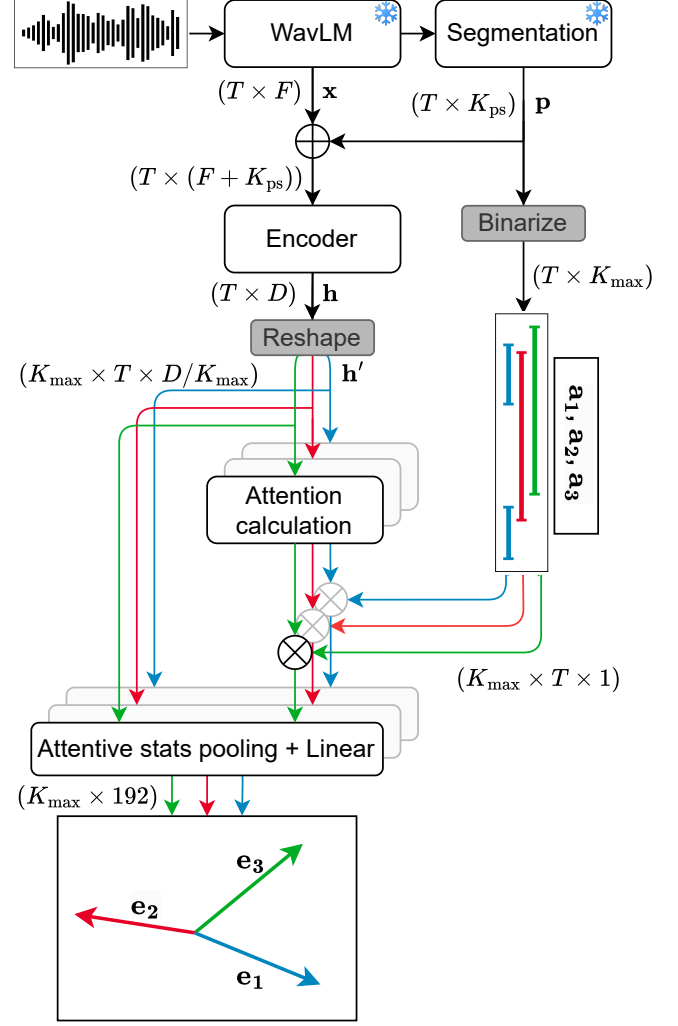


Figure 1: Proposed joint architecture for a maximum of $K_{\max} = 3$ speakers per audio chunk. We opt for a segmentation branch trained using a powerset loss, but this is not a requirement.

3. Experiments

3.1. Datasets

The feature extractor and diarization branch are trained on a composite dataset consisting of AMI-SDM [16], AliMeeting (first channel) [17], AISHELL-4 (first channel) [18], MSD-WILD [19], RAMC [20], and EGO4D [21]. Since EGO4D does not include an evaluation set, we use it only for training and validation.

The speaker embedding systems are trained on either VoxCeleb 1 and 2 utterances [7, 22] or synthetic mixtures generated from these datasets.

3.2. Data simulation

For training speaker embedding systems in multi-speaker contexts, we use 10-second synthetic VoxCeleb mixtures, following the approach in GSE. However, we modify the simulation method to better reflect real-world multi-speaker scenarios. Specifically, we allow arbitrary speaker order and permit delays

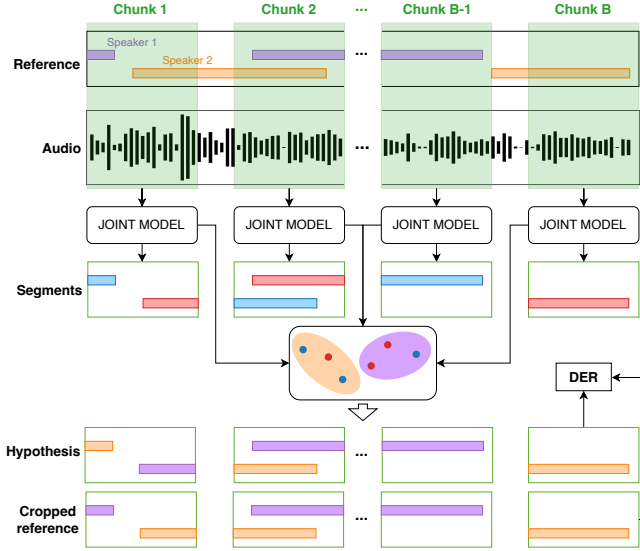


Figure 2: Proposed validation pipeline with a batch size of B .

of up to 0.5 seconds after the preceding utterance ends to introduce natural silent regions. Utterance lengths are sampled from an exponential distribution with $\lambda = 0.2$, truncated to the range $[1, 10]$ seconds. Additionally, we apply room background noise to the mixtures using data from [23] and simulated room impulse responses from [24].

3.3. Implementation details

Segmentation model. Training chunks are 10 seconds long, with a maximum of $K_{\max} = 3$ speakers per chunk. We train a standard diarization system using powerset loss, assuming that no more than two speakers are active at a time, resulting in $K_{\text{ps}} = 7$ powerset classes, assuming no more than two concurrent speakers. Our segmentation module follows the architecture from [11].

We use a WavLM Base+ model as the shared feature extractor, fine-tuned together with the segmentation module as in [10]. The learning rates are set to 10^{-5} for WavLM and 3×10^{-4} for the segmentation module, with a batch size of 32. The embedding and segmentation models use separate learnable weighted sums of the WavLM layers.

Speaker embedding model. Our speaker embedding extractor is an ECAPA-TDNN model with 1024 channels. As a baseline, we train an unguided single-embedding system on 3-second utterances with a batch size of 512.

For all other speaker embedding systems, we adopt the training strategy from [8]. We employ the Adam optimizer with a cyclical learning rate schedule over three cycles, using a batch size of 128 mixtures. This results in an effective batch size of 384 for ArcFace loss computation. Each cycle consists of 50k steps, beginning with a 1k-step warm-up phase, followed by cosine annealing decay. The learning rate starts at a peak of 10^{-3} and decays by a factor of 0.75 at the start of each new cycle.

Validation metrics. For our proposed DER-based validation metric, we randomly sample 10 files from each dataset’s validation set, yielding a total of 58 batches². The baseline validation metric is the equal error rate (EER), calculated on the widely used VoxCeleb test set 1-O, containing 37611 test trials based

²AliMeeting validation set only has 8 files

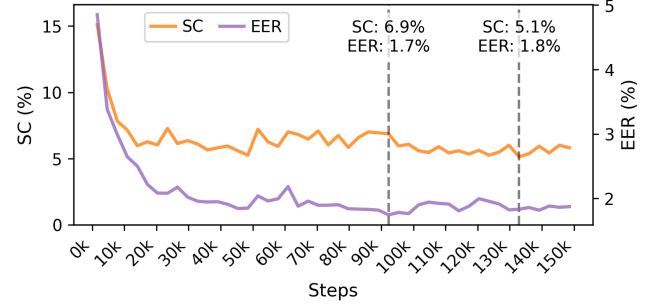


Figure 3: Validation EER and speaker confusion (SC) (%) scores as a function of step count. The optimal checkpoints based on either metric are highlighted with a dotted line.

on single-speaker utterances [7].

Speaker diarization inference. For speaker diarization inference, we use the pyannote 3.1 pipeline [25] with the same configuration as [8]. After selecting the optimal checkpoint based on the validation metric, we optimize the speaker diarization clustering hyperparameters for each system using Optuna [26]. Hyperparameter tuning is performed on the compound validation set using the multivariate Tree-structured Parzen Estimator for 100 iterations.

Evaluation. Direct evaluation of speaker embeddings in a multi-talker context would require a multi-talker real-world verification dataset, which currently does not exist. Previous work has used synthetic mixtures based on VoxCeleb to assess multi-speaker performance [8], but these do not accurately capture real-world conversational dynamics [27]. Because of this, we assess embedding quality indirectly via diarization pipeline performance. With a fixed local segmentation model, false alarm and missed detection rates are constant. Consequently, we only report speaker confusion rates, determined by clustering and directly reflecting embedding quality. Scores on the VoxCeleb test set 1-O are also reported for reference.

4. Results

A comparison of multi-speaker and single-speaker embeddings, along with different guiding methods, is presented in Table 1. All guided systems outperform the standard unguided system in diarization but underperform in EER, consistent with the findings of [8]. Switching from oracle-guided single-embedding to multi-embedding leads to a performance drop, which is expected since the encoder must now model all participating speakers rather than a single target speaker, while the model size is slightly reduced (due to scaling down the channel dimension in the attention module). However, this degradation is mitigated by replacing oracle-guided training with diarization-guided training.

Multi-speaker models are also more compact, as the channel dimension is scaled down by $K_{\max}$ after the encoder. The oracle clustering system serves as an upper bound, assuming perfect speaker clustering, with non-zero speaker confusion scores arising only from intra-chunk segmentation errors. Validation using the proposed simplified diarization pipeline demonstrates clear improvements in most cases, with comparable results for the GSE system, where the selected checkpoints had very similar performance.

Figure 3 shows the EER and speaker confusion scores as a

	Training guide	Params	Validation Metric	EER (%)	AISHELL4	AMI-SDM	AliMeeting	MSDWILD	RAMC	Macro-avg.
Oracle clustering	-	-	-	-	1.3	1.9	2.3	2.8	3.5	2.5
Single-embedding	Unguided	24.3M	EER	1.1	4.7	8.3	10.2	12.6	8.1	8.4
			DER	1.4	5.1	8.0	8.3	12.2	7.9	8.1
Single-embedding [8]	Oracle	24.3M	EER	1.8	2.9	3.8	3.6	12.1	7.3	5.9
			DER	1.9	2.7	3.7	3.6	11.8	7.2	5.9
Multi-embedding	Oracle	22.5M	EER	1.6	3.0	6.2	4.2	13.4	6.7	6.5
			DER	2.2	3.2	5.6	4.4	11.2	7.1	6.2
Multi-embedding	Diarization	22.5M	EER	1.7	3.2	3.7	5.0	11.4	7.3	6.2
			DER	1.8	2.5	3.8	3.5	11.4	7.2	5.7

Table 1: Comparison of single-speaker (single-embedding) and multi-speaker (multi-embedding) embedding systems with different guiding mechanisms and validation metrics. We report EER on VoxCeleb 1-O and speaker confusion rates on diarization datasets, as well as the macro-average (Macro-avg) for the latter. Speaker confusion using oracle clustering is included as a topline reference.

Dataset	MSE	ResNet34	SOTA
AISHELL-4	12.0	12.4	10.6 [11]
AMI-SDM	15.7	16.5	15.4 [10]
AliMeeting	15.9	17.4	13.2 [28]
MSDWILD	22.9	21.6	19.6 [11]
RAMC	11.8	11.1	11.1 [28]
Macro-average	15.0	15.1	13.4

Table 2: Comparison of DERs for multi-speaker embeddings (MSE), and ResNet34 embeddings across datasets. State-of-the-art (SOTA) DERs are provided for reference.

Dataset	Ovr. (%)	Spk. #	RTF Imp. (%)
AISHELL-4	5.0	2.0	39
AMI-SDM	14.6	2.2	43
AliMeeting	20.4	2.8	53
MSDWILD	12.4	2.0	40
RAMC	9.4	1.8	36
Macro-average	12.0	2.1	42

Table 3: Comparison of overlapping speech percentage (Ovr.), average speaker count per chunk (Spk. #), and relative RTF improvements (RTF Imp.) across datasets.

function of step count for the diarization-guided multi-speaker embedding system. The two curves display low correlation after initial fast convergence, further highlighting that the standard VoxCeleb 1-O EER is not optimal in multi-speaker applications.

In Table 2, we compare our diarization-guided multi-embedding system to a state-of-the-art ResNet-based speaker embedding model [29] from pyannote 3.1 based on DER, keeping the segmentation model the same. State-of-the-art DER scores are also provided for reference. Although the ResNet system employs a more sophisticated training strategy, including speed augmentation and large-margin fine-tuning, our system displays competitive results across the board, with significantly better results on the higher-overlap datasets AMI-SDM and AliMeeting.

Table 3 explores the real-time factor (RTF) of multi-speaker embeddings compared to single-speaker embeddings, which requires encoding each speaker in a chunk separately. We first measure the total time for inference using a diarization pipeline

with oracle speaker clustering, where no speaker embeddings are calculated. Then we measure the increase in RTF from performing clustering using either system. Comparing the results for the two systems gives us the relative decrease in RTF. We also report both the percentage of frames containing overlapped speech and the average number of speakers in a 10-second chunk sampled from the dataset. The latter directly reflects the number of separate forward passes required by the single-speaker embedding encoder, in contrast to the single pass needed for our multi-speaker approach. Even for the relatively low-overlap scenarios represented by AISHELL-4 and RAMC, the multi-embedding system achieves an RTF relative improvement of at least 36%.

4.1. Future work

Training of our diarization-guided multi-speaker embeddings relies on synthetically generated mixtures, which, while useful, fail to capture the complexity of real-world conversation dynamics [27]. Training or fine-tuning the embeddings directly on real-world data should help performance, although a comparatively small number of speakers in real-world conversational datasets poses a challenge here.

We keep the speaker embedding encoder unchanged from the single-speaker case, but since it now has to model multiple speakers, the architecture should be optimized for this.

5. Conclusion

This work introduces a novel diarization-guided training method for multi-speaker embeddings. We extend guided speaker embeddings by modeling speakers concurrently using diarization-based guidance. We also introduced a novel clustering-based validation metric for training embeddings in a multi-speaker context, which we showed to be more effective than standard speaker verification EER based on single-speaker utterances. Keeping the embedding encoder unchanged, we compared the effects of both modifications on multiple speaker diarization datasets. We showed that while switching to modeling multiple speakers concurrently degrades performance, this performance deficit is offset by switching to diarization-based guiding, which contains more information and matches testing conditions. We end up with a speaker embedding system that is smaller, more accurate, and considerably faster than comparable systems trained using previous methods.

6. Acknowledgment

This work was granted access to the HPC resources of IDRIS under the allocation AD011014274R1 made by GENCI, as well as the TalTech supercomputing resources. This work was supported by the Estonian Centre of Excellence in Artificial Intelligence (EXAI).

7. References

- [1] K. Kinoshita, M. Delcroix, and N. Tawara, “Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech,” in *Proc. Interspeech*, 2021, pp. 3565–3569.
- [2] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, “Streaming multi-talker ASR with token-level serialized output training,” in *Proc. Interspeech*, 2022, pp. 3774–3778.
- [3] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” in *Proc. ICML*, 2022, pp. 2709–2720.
- [4] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng, X. Wang, Y. Gaur, Z. Chen, J. Li, and T. Yoshioka, “Streaming speaker-attributed ASR with token-level speaker embeddings,” in *Proc. Interspeech*, 2022, pp. 521–525.
- [5] K. Kinoshita, M. Delcroix, and N. Tawara, “Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds,” in *Proc. ICASSP*, 2021, pp. 7198–7202.
- [6] N. Tawara, M. Delcroix, A. Ando, and A. Ogawa, “Ntt speaker diarization system for chime-7: Multi-domain, multi-microphone end-to-end and vector clustering diarization,” in *Proc. ICASSP*, 2024, pp. 11 281–11 285.
- [7] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “VoxCeleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, 2020.
- [8] S. Horiguchi, T. Moriya, A. Ando, T. Ashihara, H. Sato, N. Tawara, and M. Delcroix, “Guided speaker embedding,” in *Proc. ICASSP*, 2025, pp. 1–5.
- [9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [10] J. Han, F. Landini, J. Rohdin, A. Silnova, M. Diez, and L. Burget, “Leveraging self-supervised learning for speaker diarization,” in *Proc. ICASSP*, 2025, pp. 1–5.
- [11] S. Baroudi, T. Pellegrini, and H. Bredin, “Specializing self-supervised speech representations for speaker segmentation,” in *Proc. Interspeech*, 2024, pp. 3769–3773.
- [12] S. Zaiem, Y. Kemiche, T. Parcollet, S. Essid, and M. Ravanelli, “Speech self-supervised representations benchmarking: a case for larger probing heads,” *Computer Speech & Language*, vol. 89, p. 101695, 2025.
- [13] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [14] A. Plaquet and H. Bredin, “Powerset multi-class cross entropy loss for neural speaker diarization,” in *Proc. Interspeech*, 2023, pp. 1736–1740.
- [15] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” *TPAMI*, vol. 44, no. 10, pp. 5962–5979, 2022.
- [16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, “The AMI Meetings Corpus,” in *Proc. Symposium on Annotating and Measuring Meeting Behavior*, 2005.
- [17] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, “M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge,” in *Proc. ICASSP*, 2022, pp. 6167–6171.
- [18] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, “AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario,” in *Proc. Interspeech*, 2021, pp. 1736–1740.
- [19] T. Liu, S. Fan, X. Xiang, H. Song, S. Lin, J. Sun, T. Han, S. Chen, B. Yao, S. Liu, Y. Wu, Y. Qian, and K. Yu, “MSDWild: Multi-modal Speaker Diarization Dataset in the Wild,” in *Proc. Interspeech*, 2022, pp. 1476–1480.
- [20] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan, “Open source MagicData-RAMC: A rich annotated Mandarin conversational (RAMC) speech dataset,” in *Proc. Interspeech*, 2022, pp. 1736–1740.
- [21] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanov, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, “Ego4D: Around the World in 3,000 Hours of Egocentric Video,” in *Proc. CVPR*, 2022, p. 18995–19012.
- [22] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, “VoxSRC 2020: The second VoxCeleb speaker recognition challenge,” arXiv:2012.06867, 2020.
- [23] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *Proc. ICASSP*, 2017, pp. 5220–5224.
- [25] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, “pyannote.audio: neural building blocks for speaker diarization,” in *Proc. ICASSP*, 2020, pp. 7124–7128.
- [26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *Proc. KDD*, 2019, p. 2623–2631.
- [27] F. Landini, M. Diez, A. Lozano-Diez, and L. Burget, “Multi-speaker and wide-band simulated conversations as training data for end-to-end neural diarization,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [28] M. Härkönen, S. J. Broughton, and L. Samarakoon, “EEND-M2F: Masked-attention mask transformers for speaker diarization,” in *Proc. Interspeech*, 2024, pp. 37–41.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.