Speaker Diarization using multi-view contrastive learning embeddings

Rahul Dasari*
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY 14260

Abstract

Speaker diarization is a Multi-output classification problem on audio answering the question "Who spoke when?". Recently, End-to-end models have replaced traditional diarization models which use clustering-based approaches and solve issues such as dealing with overlapping speech. I have implemented an end-to-end diarization model based on the existing End-to-end Mask to Former (EEND-M2F) architecture, where in I have augmented the backbone with embeddings trained using multi-view contrastive learning. The crux of the work involved augmenting the backbone and replicating the model as described in [1] and training the model on publicly available data sets.

1 Introduction

The EEND-M2F framework reimagines speaker diarization as a segmentation problem over the temporal domain of audio. Much like object segmentation in images, where each object is identified and separated in 2D space, EEND-M2F treats the 1D audio signal as a space where each speaker constitutes a distinct "object" to be detected and segmented over time. The model outputs a sequence of binary masks—one per speaker—indicating the presence or absence of each speaker at each time frame. This formulation directly addresses the core diarization question: who spoke when, and does so without requiring traditional clustering or heuristic-based steps.

This project aims to improve speaker discrimination within this framework by enhancing the encoder backbone's representation capacity. To do this, I integrate speaker-aware embeddings derived through multi-view contrastive learning, a self-supervised technique that learns to project different augmentations or "views" of the same speaker's audio into a shared embedding space, while pushing apart representations of different speakers.

These embeddings are trained separately using speaker recognition models where each view corresponds to a distinct transformation of a given speech segment. The model learns to recognize that these transformed views belong to the same speaker, while maintaining separation from others. This multi-view setting helps capture more robust, invariant speaker characteristics.

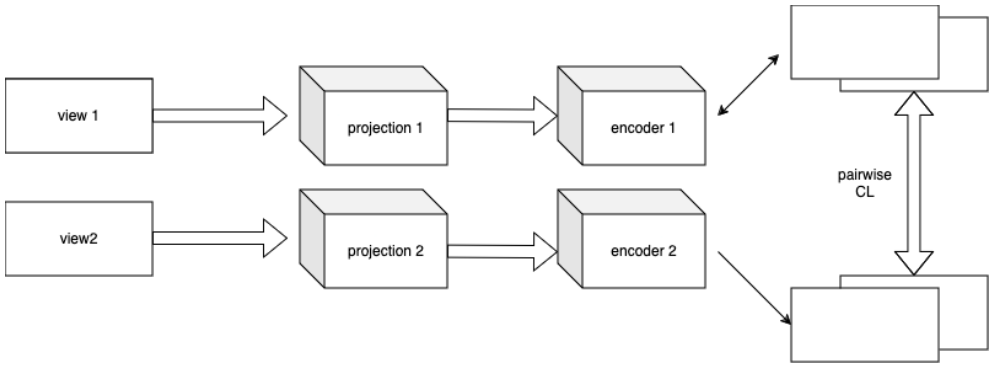
Once trained, these embeddings are incorporated into the EEND-M2F model using an early fusion strategy—that is, they are concatenated or combined with the input features before being passed into the mask module of the diarization model. This early integration allows the model to use both raw acoustic cues and speaker-discriminative information simultaneously during learning. The hypothesis is that this additional speaker-aware information will help the improve the model's accuracy.

*This project is in fulfillment of credits for the masters project

1.1 Multiview-PairwiseCL

Multiview Pairwise Contrastive Learning (Multiview-PairwiseCL) is a self-supervised learning framework designed to learn discriminative and robust feature representations in the absence of sufficient annotated data. Unlike traditional supervised approaches that rely heavily on labeled datasets, Multiview-PairwiseCL operates by constructing positive and negative sample pairs from multiple augmented views of the same input data, enabling the model to learn invariances and semantic structures inherent in the data distribution.

Formally, given an input audio segment, multiple augmentations are applied to generate different "views" of the same underlying content. For each anchor view, a corresponding positive view is selected from the same source segment, while negative views are drawn from different segments within a batch. The objective is to minimize the distance between embeddings of positive pairs while maximizing the distance between embeddings of negative pairs.



In my framework, I leverage Multiview Pairwise Contrastive Learning (Multiview-PairwiseCL) to learn high-quality, speaker-specific latent representations from speech data within a supervised training regime. Unlike its typical application in self-supervised contexts, I adapt Multiview-PairwiseCL to utilize ground-truth speaker labels to explicitly construct positive and negative sample pairs. Specifically, speech segments originating from the same speaker are treated as positive pairs, while segments from different speakers form negative pairs. These curated pairwise relationships guide the model to optimize a contrastive objective that minimizes the embedding distance between utterances of the same speaker while maximizing the separation between those of different speakers in the latent representation space. By integrating supervision into the contrastive learning pipeline, I ensure that the learned representations are not only invariant to superficial acoustic variations but are also highly discriminative with respect to speaker identity—thus enabling more effective downstream tasks.

1.2 EEND-M2F

My work builds on EEND-M2F, a state-of-the-art neural diarization architecture that generalizes the Mask2Former framework, originally proposed for semantic segmentation in the image domain to one-dimensional temporal audio sequences. Recognizing the domain agnostic design of Mask2Former, EEND-M2F repurposes these capabilities for speaker diarization.

In my implementation, I adhere to the architectural principles described in [1], with modifications to accommodate the model on the gpus to which I have access.

The core of my system is built on the integration of Multiview PairwiseCL and EEND-M2F. The multiview pairwiseCL embeddings are then fused into EEND-M2F, which extends advanced segmentation strategies to the audio domain. Together, these components form a cohesive framework for scalable, accurate, and interpretable speaker diarization.

2 Method

I generated the embeddings using a multi-view contrastive self-supervised learning (SSL) pre-training technique, Pairwise-CL. Pairwise-CL leverages the NT-Xent loss to align representations across

diverse audio views, fostering invariant and discriminative embeddings. The approach incorporates pre-trained Resnet TDNN and ECAPA-TDNN as initial views.

The embeddings are produced using two speaker recognition systems. The first system is a ResNet-TDNN model, which combines residual network layers with a time-delay neural network (TDNN) [4]. It is trained with Additive Margin Softmax Loss, and speaker verification is performed by calculating cosine distance between the extracted embeddings. The second system is an ECAPA-TDNN model, which integrates convolutional and residual blocks, using attentive statistical pooling for embedding extraction. This model is also trained using Additive Margin Softmax Loss and employs cosine distance for speaker verification.

Speaker embeddings in my system are extracted using two state-of-the-art speaker recognition architectures. The first model is a hybrid ResNet-TDNN architecture, which integrates residual convolutional layers for deep feature extraction with time-delay neural network (TDNN) layers to capture temporal dependencies in the speech signal [4]. This model is optimized using the Additive Margin Softmax (AM-Softmax) loss, which enhances inter-class separability in the embedding space. For speaker verification, the similarity between embeddings is quantified using cosine distance.

Pairwise-CL aligns the embeddings from these two models, by combining complementary features and enforcing alignment across views, the proposed methodology ensures the embeddings are invariant to noise and discriminative for speaker-specific characteristics.

The Loss function is computed according to the following equations:

$$l_i^{l \to j} = -\log \frac{\delta(z_i^l, z_j^l)}{\sum_{k=1}^{N} \delta(z_i^l, z_k^l)}, \tag{1}$$

where $\delta(z_i^l, z_j^l)$ is the exponent of a similarity function. The total loss is:

$$L^{l \to j} = \frac{1}{N} \sum_{i=1}^{N} \left(l_i^{l \to j} + l_j^{l \to i} \right) \tag{2}$$

We compute losses between all pairs of views:

$$\mathcal{L} = \sum_{1 \le k < k' \le K} \left(L^{k \to k'} + L^{k' \to k} \right) \tag{3}$$

The proposed loss function aims to maximize the similarities for multi-view representations for to the same instance.

After this pretraining, I use these embeddings in the encoder backbone of the EEND-M2F and further train the model for the speaker diarization downstream task.

2.1 Encoder Backbone

The input sequence X gets downsampled via convolutional layers to a 1/10th the resolution, after which it passes through Conformer layers to produce the low-resolution latent sequence.

$$\mathcal{E} = \text{Conformer}(\text{ConvDown}(X)) \tag{4}$$

2.2 Mask module

The mask module combines the latent space embedding with queries to generate probabilities for each query.

$$\text{MaskModule}(Q, \mathcal{E}) := \mathcal{E} \cdot \text{MLP}(Q)^T \tag{5}$$

$$\tilde{Y} = \sigma(\text{MaskModule}(Q, \mathcal{E})) \tag{6}$$

2.3 Query Module

The query module processes a set of input queries $Q^{(\ell)}$ alongside the latent acoustic representation $\mathcal{E}$, producing an updated set of queries $Q^{(\ell+1)}$ that maintains the original shape of $Q^{(\ell)}$. Initially, masked cross-attention is computed between $Q^{(\ell)}$ and $\mathcal{E}$, where the attention mask is derived from intermediate diarization logits defined as $M^{(\ell)} = \text{MaskModule}(Q^{(\ell)}, \mathcal{E})$. The resulting masked attention outputs are then propagated through a stack of Transformer blocks comprising multi-head self-attention and feed-forward sublayers, ultimately yielding the refined queries. Formally:

$$Q' = \text{LN}(\text{MHA}(Q^{(\ell)} + P, \mathcal{E}, \mathcal{E}; M^{(\ell)}) + Q^{(\ell)}) \tag{7}$$

$$Q'' = \text{LN}(\text{MHA}(Q' + P, Q' + P, Q') + Q') \tag{8}$$

$$Q^{(\ell+1)} = \text{LN}(\text{FF}(Q'') + Q'') \tag{9}$$

2.4 Query classification module

I employed a simple classification layer to decide which columns to from $\tilde{Y}$.

$$\hat{\tilde{p}} = \sigma(\text{Linear}(Q)) \tag{10}$$

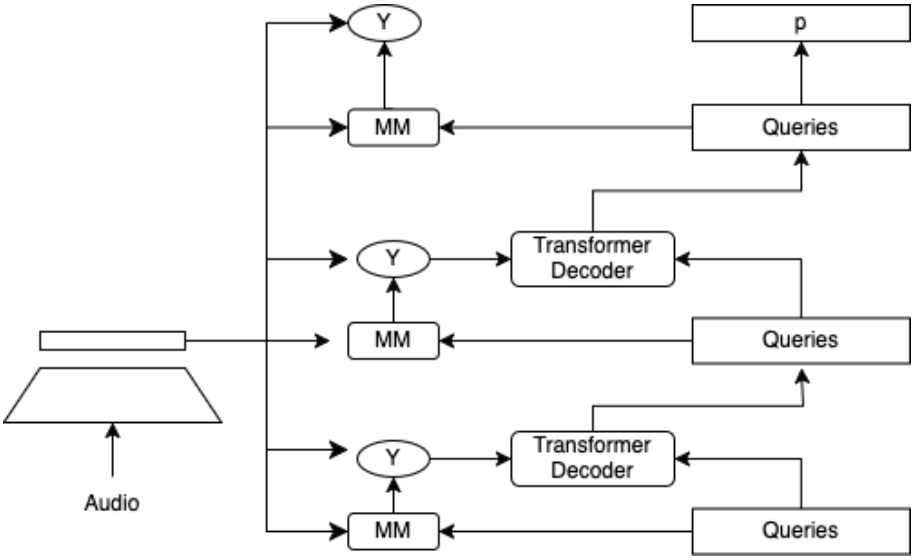


Figure 1: The Model

2.5 Loss function

I employed Permutation invariant training, wrapping Binary cross entropy loss for $\tilde{p}$, Binary cross entropy loss for $\tilde{Y}$ and Dice loss. The loss is the sum of these three losses during training. Permutation invariant training uses a hungarian matching step to find the optimal permuation of predictions that minimizes the Binary cross entropy loss of $\tilde{Y}$.

3 Results

The training of the embeddings using Pairwise-CL yielded promising results. As depicted in the Figure below, the loss recorded during the training phase shows a steady decline over 200 iterations for all three datasets. This consistent reduction in loss indicates that the network is effectively learning to cluster embeddings based on speaker similarity, a critical step toward robust speaker diarization.

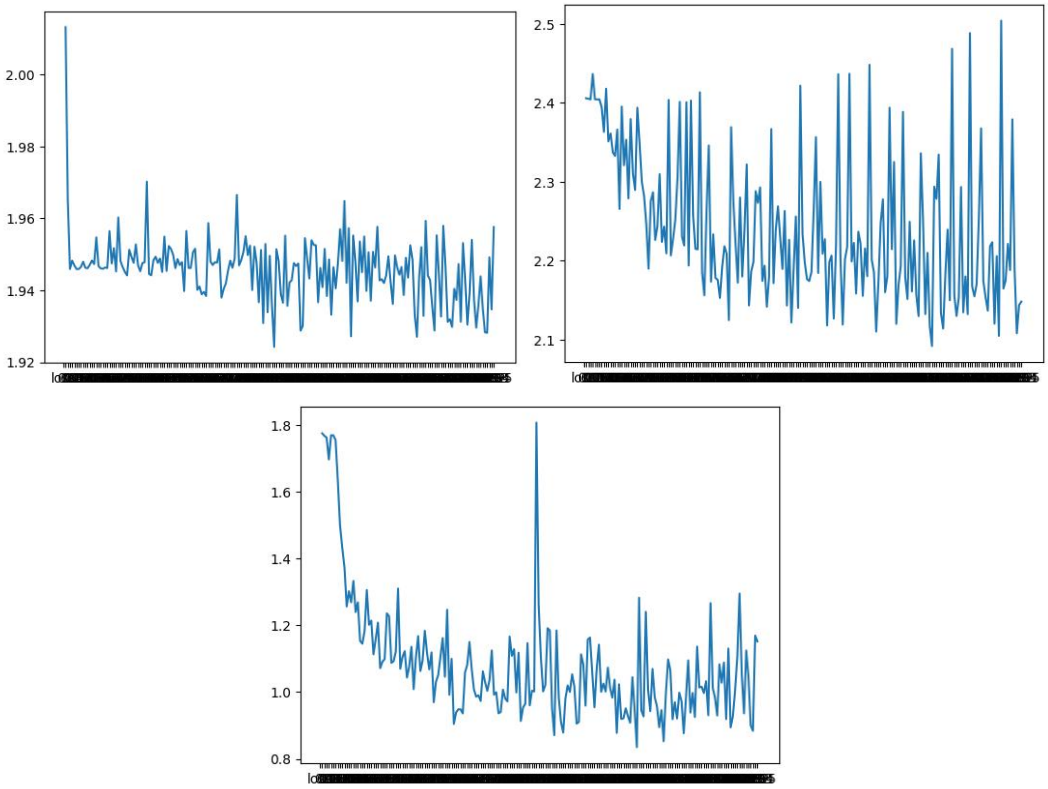


Figure 2: Contrastive Learning loss on datasets AMI, ICSI, and Voxconverse

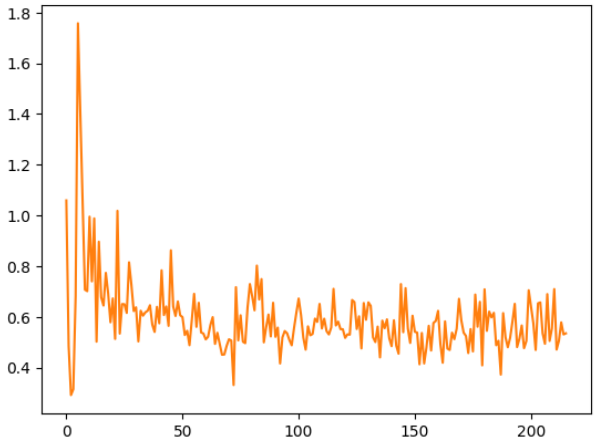


Figure 3: Model Loss

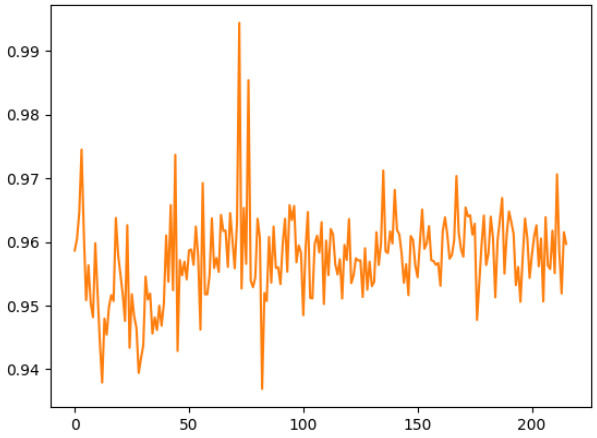


Figure 4: Model Accuracy

I trained the model using the aforementioned architectural principles. I added 6 layers of transformer decoders, as mentioned in [1]. I compressed the raw audio by a factor of 100. The results presented below pertain to this model configuration. The plots show the average loss and average accuracy per sample. We see a steady increase in accuracy and decrease in loss for 50 training samples after which this model plateaus. I believe this might be caused by the severe compression of the raw audio and the query dimensions in the transformer decoder layers.

References

[1] M. Härkönen, S. J. Broughton, and L. Samarakoon, "EEND-M2F: Masked-attention mask transformers for speaker diarization," *arXiv*, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.12600.

[2] B. Khaertdinov, P. Jeuris, A. Sousa, and E. Hortal, "Exploring Self-Supervised Multi-view Contrastive Learning for Speech Emotion Recognition with Limited Annotations," *arXiv*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.07900.

[3] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," *arXiv*, 2020. [Online]. Available: https://arxiv.org/pdf/2112.01527.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *arXiv*, May 2020. [Online]. Available: https://arxiv.org/pdf/2005.07143.

[5] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," *arXiv*, Jun. 2021. [Online]. Available: https://arxiv.org/pdf/2106.04624.

[6] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, and P. A. Torres-Carrasquillo, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020. [Online]. Available: https://doi.org/10.1016/j.csl.2019.101026.

[7] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, Oct. 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6518171.