# MULTI-VIEW SPEAKER EMBEDDING LEARNING FOR ENHANCED STABILITY AND DISCRIMINABILITY

*Liang He*$^{1,2,3,\dagger}$ *Zhihua Fang*$^{1,2}$ *Zuoer Chen*$^{3}$ *Minqiang Xu*$^{4}$ *Ying Meng*$^{1,2}$ *Penghao Wang*$^{3}$

$^{1}$School of Information Science and Engineering, Xinjiang University, Urumqi, China
$^{2}$Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi, China
$^{3}$Department of Electronic Engineering, Tsinghua University, Beijing, China
$^{4}$iFly Digital Technology, Hefei, China

## ABSTRACT

Deep neural network models based on x-vector have become the most popular framework for speaker recognition, and the quality of speaker features (embedding) is important for open-set tasks such as speaker verification and speaker diarization. Currently, the most popular loss function is based on margin penalty, however, it only considers enlarging the inter-class distance while neglecting to reduce the intra-class feature differences. Therefore, we propose a multi-view learning approach that divides the training process into two views from the speaker embedding level. The classification view focuses on distinguishing the discriminability of different speakers, while the clustering view focuses on shrinking the feature boundaries of the same speaker, making intra-class differences smaller. The combined effect of the two perspectives achieves large inter-class distances and small intra-class distances, resulting in the extraction of more discriminative and stable speaker embedding. We test the performance of the method on both speaker verification and speaker diarization tasks, and the results demonstrate the effectiveness of our approach.

***Index Terms*—** Speaker embedding, speaker verification, speaker diarization, deep clustering

## 1. INTRODUCTION

Speaker recognition is a biometric technology, which is divided into the following research directions: speaker identification is to determine which of the registered speakers a speech belongs to, speaker verification is to determine whether both ends of the speech belong to the same speaker, and speaker diarization is to solve the problem of ”who spoke when”. The first one belongs to the classification task, while the last two belong to the open-set task, which needs to extract speaker's features (embedding) for comparison, and the quality of speaker's embedding extracted by the model determines the performance of these tasks.

During the past several years, x-vector [1] has become the dominant framework in the field of speaker recognition because of its simplicity and excellent performance, and most of the technologies have been developed around x-vector [2, 3, 4]. A typical x-vector framework consists of a neural network backbone, a pooling layer for extracting frame-level features, and several fully-connected layers (FC). The output of the penultimate full-connected layer is considered to be speaker features (embedding), which are compact representations of the speaker's identity and are used for open-set recognition tasks such as speaker verification or speaker diarization.

Drawing on the experience of face recognition, early researches used the cross-entropy loss function (CE) as an objective function for speaker recognition [5]. CE works to distinguish different speaker embeddings, but does not take into account the fact that embeddings located at decision boundary may be more similar to neighboring speaker embeddings, which is a fatal hazard for open-set tasks such as speaker verification. The most effective approach makes the inter-class distance between speakers larger and the intra-class distance smaller when training the classification model. Therefore various margin-based objective functions (e.g., A-Softmax [6], AM-Softmax [7], AAM-Softmax [8], etc.) have been proposed, and although these max-margin loss functions effectively separate the target speakers from the non-target speakers, they do not take into account how the embedding belonging to the same speaker can be more compact. Especially for tasks such as speaker diarization, which requires clustering of speech segments, compact and stable speaker embedding is a core requirement.

Inspired by deep clustering [9], we proposed a **M**ulti-**V**iew **S**peaker **E**mbedding (**MVSE**) learning method. We divide the training process from the speaker embedding level into two views: the classification view uses max-margin based objective function, which is dedicated to distinguishing between target and non-target speakers, and the clustering view uses clustering method to cluster the embeddings, which

This research supported by National Key R&D Program of China (2022ZD0115801), National Natural Science Foundation of China No. 62366051.

$^{\dagger}$ Corresponding author (heliang@mail.tsinghua.edu.cn).

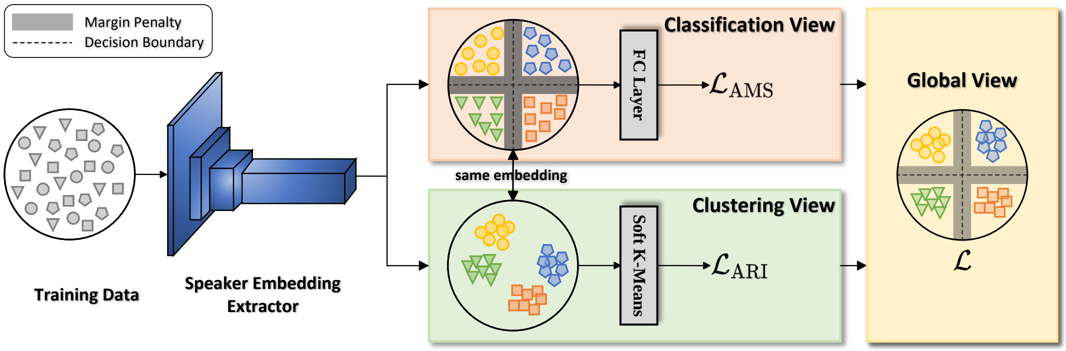

**Fig. 1**. **The Multi-View Speaker Embedding Learning Framework. Classification view** aims to enlarge the distance between the embedding of different speakers. **Clustering view** aims to shrink the boundaries of embedding regions belonging to the same speaker. **Global view** is equivalent to the combination of the above two views, which can simultaneously enlarge the inter-class distance and reduce the intra-class distance.

is dedicated to shrinking the boundaries of speaker clusters. The combination of the two views allows the speaker model to extract embeddings that are both distinguishable and stable. The contributions of this paper are as follows:

- We proposed a multi-view learning approach to enhance the quality of speaker embedding by dividing the classification view and clustering view at the embedding level, and the two views jointly supervise the speaker model training;
- In order to efficiently assess the quality of the extracted embedding, we conducted experiments on both speaker verification and speaker diarization tasks, and the results demonstrated the effectiveness of the multi-view learning.

## 2. METHODS

In this section, we describe the methods used for the classification view and the clustering view, as well as the overall framework, and then introduce our proposed multi-view learning method.

### 2.1. Classification View

The famous Max-Margin based loss functions are A-Softmax [6], AM-Softmax [7], AAM-Softmax [8], etc., among which AM-Softmax becomes one of the popular objective functions due to its excellent performance and easy calculation, and its formula is as follows:

$$\mathcal{L}_{\text{AMS}} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s\cdot(\cos(\theta_{y_i})-m)}}{e^{s\cdot(\cos(\theta_{y_i})-m)} + \sum_{j=1,j\neq y_i}^{C} e^{s\cdot\cos(\theta_j)}}, \tag{1}$$

where $N$ is the batch size and $C$ is the number of speakers. $\cos(\theta_j) = W_j^T \cdot \boldsymbol{x}_i$, $y_i$ denotes the ground-truth label, $\boldsymbol{x}_i$ is the embedding of the i-th speech, and $W$ is the weight of the last FC layer, which can also be regarded as each speaker's embedding prototype.

By applying the margin penalty to $\cos(\theta_{y_i})$, it can make $\boldsymbol{x}_i$ far away from the decision boundary, as shown in Fig. 1, there will be a blank space retained between the embedding of the different speakers, and in this way, achieving the discriminability between the target and non-target speakers.

### 2.2. Clustering View

Inspired by deep clustering network (DCN) [9], combining DNN and K-Means can make potential representations (embedding) exhibit strong cohesion, so we introduce clustering methods on segment-level embedding. The discreteness of traditional K-Means would cause gradient propagation blocking at the training stage [10]. To solve it, we adopt a soft K-Means to make it differentiable and calculate backward propagation. Given a batch of inputs $\boldsymbol{f}_i = f(\boldsymbol{x}_i), 1 \le i \le N$, we introduce an auxiliary $\boldsymbol{z}_i$, which is

$$\boldsymbol{z}_{i,j} = \frac{\exp(-\kappa\|\boldsymbol{f}_i - M_j\|^2)}{\sum_{k=1}^{K}\exp(-\kappa\|\boldsymbol{f}_i - M_k\|^2)}, \tag{2}$$

where $M_j$ is the $j$-th column vector of $M$, $K$ is the number of centroids and $\kappa$ is a hyper parameter. And we try to minimize the $\sum_{i=1}^{N}\|f(\boldsymbol{x}_i) - M\boldsymbol{z}_i\|^2$ by the Expectation-Maximization (EM) algorithm. And the optimization procedure is given in the Algorithm 1.

After the soft K-Means clustering, we calculate the mismatch between $N$ predicted clustering labels $z_i$ and $N$ ground-truth labels $y_i$, and compute the ARI loss [11] as follows:

$$\mathcal{L}_{\text{ARI}} = \frac{-2(N_1N_4 - N_2N_3)}{(N_1+N_2)(N_3+N_4)+(N_1+N_3)(N_2+N_4)}, \tag{3}$$

```
Algorithm 1: Soft K-Means Algorithm
  Input: A batch of embeddings $F = \{\boldsymbol{f}_i\}$, Number
         of batch size $N$, Number of centroids $K$,
         Hyper parameter $\kappa$, Loop number $LP$
  repeat
    E-step
      Compute auxiliary $\boldsymbol{z}_i$, $1 \leq i \leq N$:
        $\boldsymbol{z}_{i,j} = \frac{\exp(-\kappa\|\boldsymbol{f}_i - M\boldsymbol{i}_j\|^2)}{\sum_{k=1}^{K}\exp(-\kappa\|\boldsymbol{f}_i - M\boldsymbol{i}_k\|^2)}$
    M-step
      Update centroids $M$:
        $M_k \leftarrow M_k + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_{i,k}\boldsymbol{f}_i$
  until convergence;
  Output: Centroids $M$
```

in which:

$$
\begin{array}{l}
N_1 = \sum_{i=1}^{N}\sum_{i'=i+1}^{N} I(y_i \neq y_{i'})d_{i,i'}, \\
N_2 = \sum_{i=1}^{N}\sum_{i'=i+1}^{N} I(y_i \neq y_{i'})(1 - d_{i,i'}), \\
N_3 = \sum_{i=1}^{N}\sum_{i'=i+1}^{N} I(y_i = y_{i'})d_{i,i'}, \\
N_4 = \sum_{i=1}^{N}\sum_{i'=i+1}^{N} I(y_i = y_{i'})(1 - d_{i,i'}),
\end{array}
\tag{4}
$$

where $I(\cdot)$ is the indicator function, i.e., $I(A) = 1$ if $A$ is true and 0 otherwise. $d_{i,i'} = \frac{1}{2}\sum_{k=1}^{K}|z_{i,k} - z_{i',k}|$ is the total variation distance [11] between $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i'}$, and $z_{i,k}$ is the soft predicted clustering label for one segment of K clusters. By computing the total mismatch distance of the predicted clustering labels and the ground-truth labels under traversal, we can calculate the ARI loss as the clustering loss to enhance the cohesion of embeddings from the same class.

### 2.3. Global View

The global view is a combination of the classification view and the clustering view. The overall objective function is as follows:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{AMS}} + (1 - \lambda)\mathcal{L}_{\text{ARI}}, \tag{5}$$

where $\lambda$ is the parameter that controls the weights of classification loss and clustering loss. For classification view we choose AM-Softmax loss and for clustering view we choose ARI loss. In fact, any kind of max-margin based loss can be used in the classification view, and the same is true for the clustering view.

## 3. EXPERIMENTS

### 3.1. Datasets

We trained the speaker embedding extractor using the development set of VoxCeleb2 [12], which consists of 5994 speakers and more than 1 million utterances. To evaluate the quality of speaker embedding extracted by our method, we tested it on both speaker verification and speaker diarization tasks.

For the speaker verification task, we test on the VoxCeleb1 [13] and VoxMovies [14] datasets. VoxCeleb1 consists of 1,211 speakers and 148,642 utterances, which contains the three test sets Vox-O, Vox-E, and Vox-H. VoxMovies consists of 856 speakers and 8,905 utterances, which includes the 5 evaluation sets (E1-E5). For the speaker diarization task, we tested on the VoxConverse [15] dataset. The development set contains of 216 multi-speaker videos covering 1,218 minutes with 8,268 speakers turned annotate. The test set contains approximately 232 videos covering 2,612 minutes.

### 3.2. Implementation Details

We used thin-ResNet backbone, statistic pooling and AM-Softmax loss [7] for training, and augmented the data 5 times using MUSAN [18] and RIRs [19]. A 64-dimensional FBank is used as features, and a fixed 2-second speech segment is acquired by VAD as model input. During the training phase, we use the SGD optimizer with 0.9 momentum and $10^{-4}$ weight decay. Linear warmup and step warmup strategies are used in training with a $10^{-1}$ initial learning rate and a $5 \times 10^{-4}$ final learning rate. The mini-batch size is set to 128.

For the speaker verification task, equal error rate (EER) and minimum detection cost function (minDCF) are used as evaluation metrics. For the speaker diarization task, the diarization error rate (DER) and jaccard error rate (JER) are used as evaluation metrics.

### 3.3. Results on the speaker verification task

As shown in Table 1, the first line (thin-ResNet34 + AM-Softmax) and the third line (thin-ResNet34 + AM-Softmax) have the same network and loss functions. The EERs of the third line tested on VoxCeleb1-O, VoxCeleb1-E and VoxCeleb1-H databases demonstrate the effective training methods of our system. We regard the third line as our baseline to verify the MVSE. The fourth line (thin-ResNet34 + MVSE) has the same network but different loss functions with the baseline. The EERs of thin-ResNet34 with the MVSE are $0.95\%$ on the VoxCeleb1-O, $1.06\%$ on the VoxCeleb1-E and $1.88\%$ on the VoxCeleb1-H database, relatively reduced by $8.7\%$, $4.5\%$ and $10.9\%$, respectively, compared with the AM-Softmax loss, which demonstrates significant effectiveness of our proposed the MVSE. The minDCFs on these three datasets also reduced by $11.3\%$, $5.6\%$ and $9.9\%$, respectively.

**Table 1**. *Speaker verification results on the VoxCeleb1 dataset.*

| Network | Loss | VoxCeleb1-O | | VoxCeleb1-E | | VoxCeleb1-H | |
|---|---|---|---|---|---|---|---|
| | | **EER** | minDCF$_{0.01}$ | **EER** | minDCF$_{0.01}$ | **EER** | minDCF$_{0.01}$ |
| thin-ResNet34[16] | AM-Softmax loss | 3.23 | - | 3.13 | - | 5.06 | - |
| thin-ResNet34[17] | Adaptive margin circle loss | 1.44 | 0.161 | 1.58 | 0.170 | 2.64 | 0.240 |
| thin-ResNet34 | AM-Softmax loss | 1.04 | 0.106 | 1.11 | 0.126 | 2.11 | 0.203 |
| thin-ResNet34 | **MVSE** | **0.95** | **0.094** | **1.06** | **0.119** | **1.88** | **0.183** |
| thin-ResNet50 | **MVSE** | 1.13 | 0.131 | 1.22 | 0.131 | 2.15 | 0.201 |
| thin-ResNet101 | **MVSE** | 1.00 | 0.101 | 1.17 | 0.127 | 2.05 | 0.190 |

**Table 2**. *Speaker verification results on the VoxMovies dataset.*

| EER | E-1 | E-2 | E-3 | E-4 | E-5 |
|---|---|---|---|---|---|
| Angular loss [14] | 9.72 | 10.58 | 10.58 | 12.52 | 14.11 |
| AM-Softmax loss | 5.30 | 6.83 | 7.57 | **7.41** | 9.65 |
| **MVSE** | **5.05** | **6.66** | **7.06** | 7.47 | **9.37** |

The second line is the recently proposed adaptive margin circle (AMC) loss [17]. Under the same network configuration, the MVSE is better than the AMC loss in terms of EER and minDCF with a large margin. By comparing the fourth, fifth and sixth lines, we notice that the deeper network structure does not bring gains. In some cases, the performance has deteriorated. We attribute the improvements mainly to the introduction of clustering loss, which effectively increases the cohesiveness of embeddings from the same speaker.

Furthermore, we verify the MVSE on the VoxMovies dataset, which is more challenging for its domain mismatch. From the Table 2, we can see that the thin-ResNet34 with MVSE reduces the EER by 48.0% and minDCF by 24.0%, in average, under five conditions (E-1 to E-5), compared with the reported results [14], and achieves a 6.5% reduction in EER, 6.8% reduction in minDCF compared with the baseline system on the VoxMovies dataset, which means that the MVSE is robust under variant emotion, accents and backgrounds, and performs well in the case of domain mismatch.

### 3.4. Results on the speaker diarization task

Since the VoxConverse dataset is also the official dataset of VoxSRC-20 Track 4, we choose some excellent results with only AHC clustering algorithm reported in this competition for comparison [20]. As shown in Table 3, the first line to the third line are the results reported in the competition and the next two lines are our MVSE diarization systems consisting of the embedding extractor and the same AHC back-end clustering algorithm. Our baseline system (thin-ResNet34 + AM-Softmax) in the fourth line does not perform well on the DER and JER metrics on the VoxConverse dataset compared with other systems. However, the overall results of our MVSE diarization system on the speaker diarization task with no overlapped speech detection are the best. We achieve an 8.4% and 4.9% DER reduction and an 11.1% and 7.1% JER reduction on VoxConverse Dev and Test datasets compared with the baseline system using the same network but different loss functions, which proves the MVSE extracts more discriminative embeddings for speaker diarization task.

**Table 3**. *Comparison with the speaker diarization results reported in the VoxSRC-20 competition on VoxConverse dataset.*

| System | VoxConverse-Dev | | VoxConverse-Test | |
|---|---|---|---|---|
| | **DER** | **JER** | **DER** | **JER** |
| Baseline[15] | 7.7 | - | 21.75 | 51.89 |
| mandalorian[21] | 4.91 | 19.90 | 8.54 | 20.58 |
| landini[22] | 5.73 | 21.56 | 8.12 | 18.35 |
| AM-Softmax loss | 5.11 | 14.79 | 5.46 | 18.61 |
| **MVSE** | **4.68** | **13.15** | **5.19** | **17.28** |

The motivation of MVSE originates from getting clustering-friendly embeddings, which is the core issue of speaker diarization task. By integrating neural network and clustering, we make the embeddings of the same speaker to be better aggregated, while separating embeddings from different speakers at the same time. So, we can extract more discriminant embeddings with the help of MVSE to accomplish the speaker diarization task.

## 4. CONCLUSION

We proposed a multi-view speaker embedding learning method to train speaker models and make the extracted embedding more robust to different tasks. The classification view focuses on improving the discriminability of speaker embedding and the clustering view focuses on improving the cohesiveness of speaker embedding. Our experimental results on VoxCeleb1, VoxMovies, and VoxConverse show that MVSE achieves excellent performance on both speaker verification and speaker diarization tasks, reflecting the discriminability and cohesiveness of multi-view speaker embedding.

## 5. REFERENCES

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP*, 2018, pp. 5329–5333.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. of INTERSPEECH*, 2020, pp. 3830–3834.

[3] Zhenduo Zhao, Zhuo Li, Wenchao Wang, and Pengyuan Zhang, "Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification," in *Proc. of ICASSP*, 2023, pp. 1–5.

[4] Jiadi Yao, Chengdong Liang, Zhendong Peng, Binbin Zhang, and Xiao-Lei Zhang, "Branch-ECAPA-TDNN: A Parallel Branch Architecture to Capture Local and Global Features for Speaker Verification," in *Proc. of INTERSPEECH*, 2023, pp. 1943–1947.

[5] Zhongxin Bai and Xiao-Lei Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[6] Yutian Li, Feng Gao, Zhijian Ou, and Jiasong Sun, "Angular softmax loss for end-to-end speaker verification," in *Proc. of ISCSLP*, 2018, pp. 190–194.

[7] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[8] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2022.

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. of ECCV*, 2018, pp. 139–156.

[10] Xuesong Yin, Songcan Chen, and Enliang Hu, "Regularized soft k-means for discriminant analysis," *Neurocomputing*, vol. 103, pp. 29–42, 2013.

[11] Tomoharu Iwata, "Meta-learning representations for clustering with infinite gaussian mixture models," *Neurocomputing*, vol. 549, pp. 126423, 2023.

[12] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of INTERSPEECH*, 2018, pp. 1086–1090.

[13] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. of INTERSPEECH*, 2017, pp. 2616–2620.

[14] Andrew Brown, Jaesung Huh, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Playing a part: Speaker verification at the movies," in *Proc. of ICASSP*. IEEE, 2021, pp. 6174–6178.

[15] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Proc. of INTERSPEECH*, 2020, pp. 299–303.

[16] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. of ICASSP*. IEEE, 2019, pp. 5791–5795.

[17] Runqiu Xiao, Xiaoxiao Miao, Wenchao Wang, Pengyuan Zhang, Bin Cai, and Liuping Luo, "Adaptive Margin Circle Loss for Speaker Verification," in *Proc. of INTERSPEECH*, 2021, pp. 4618–4622.

[18] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[19] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. of ICASSP*, 2017, pp. 5220–5224.

[20] Arsha Nagrani, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[21] Xiong Xiao, Naoyuki Kanda, Zhuo Chen, Tianyan Zhou, Takuya Yoshioka, Sanyuan Chen, Yong Zhao, Gang Liu, Yu Wu, Jian Wu, et al., "Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020," in *Proc. of ICASSP*. IEEE, 2021, pp. 5824–5828.

[22] Federico Landini, Ondřej Glembek, Pavel Matějka, Johan Rohdin, Lukáš Burget, Mireia Diez, and Anna Silnova, "Analysis of the but diarization system for voxconverse challenge," in *Proc. of ICASSP*, 2021, pp. 5819–5823.