Introducing Discriminative Speaker Embeddings for Voice Timbre Attribute Detection ⋆

Zhida Song[1] and Liang He[1,2,3,4]

[1] School of Computer Science and Technology, School of Intelligence Science and Technology, Xinjiang University, Urumqi 830017, China
[2] Xinjiang Multimodal Information Technology Engineering Research Center, Urumqi 830017, China
[3] Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830017, China
[4] Department of Electronic Engineering, and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
szd@stu.xju.edu.cn, heliang@mail.tsinghua.edu.cn

Abstract. This paper presents the AuroraLab system for the Voice Timbre Attribute Detection (vTAD) 2025 Challenge. In this challenge, we propose a novel framework that introduces discriminative speaker embeddings (DSE) into the vTAD task, termed DSE-vTAD. DSE-vTAD leverages strong speaker embedding extractors to obtain discriminative speaker embeddings. In addition, unlike the challenge's baseline system, DSE-vTAD concatenates a pair of speaker embeddings along with their Hadamard product features. Compared with the baseline system, DSE-vTAD achieves significant performance improvements. On the unseen test set, the best DSE-vTAD system achieves 91.31% Avg ACC and 8.54% Avg EER. On the seen-speaker test set, the best DSE-vTAD system achieves 97.32% Avg ACC and 2.72% Avg EER.

Keywords: Voice timbre attribute detection · vTAD 2025 Challenge · Speaker embedding · Hadamard product.

1 Introduction

Voice timbre attribute detection (vTAD) [1, 2] is a task that aims to determine whether there exists a relative strength difference in a specific timbre attribute between two speech utterances. Timbre attributes refer to perceptual descriptors used by listeners to characterize a speaker's timbre based on auditory cues, such as bright, coarse, round, and magnetic. This task aims to deepen the understanding of voice timbre by analyzing and modeling speaker-specific timbre attributes, thereby advancing the development of timbre-related speech technologies such as explainable speaker recognition [3] and speaker generation [4–6].

⋆ This work was supported by the National Natural Science Foundation of China under Grant 62366051.
Corresponding Author: Liang He.

In recent years, speaker-related tasks have commonly employed low-dimensional speaker embeddings to represent target speakers [7,8]. This has motivated extensive research on extracting discriminative speaker embeddings, including using larger-scale datasets [9,10], improvements in network architectures [11–17], and the disentanglement of speaker-independent information [18–22].

In this paper, we propose a novel framework for the vTAD task by introducing discriminative speaker embeddings (DSE) called DSE-vTAD. First, DSE-vTAD employs a pretrained speaker embedding extractor to derive speaker embeddings from an ordered pair of speech utterances. Then, unlike the challenge's baseline system[1], DSE-vTAD concatenates the two speaker embeddings along with their Hadamard product to explicitly capture the dimension-wise relationships between them. Finally, a classification network performs voice timbre attribute detection based on the concatenated features. Specifically, we evaluate four speaker embedding extractors within the DSE-vTAD framework: ECAPA-TDNN [15], FACodec [21], SimAM-ResNet34 [13], and SimAM-ResNet100 [13]. On the vTAD 2025 Challenge test sets, the DSE-vTAD system with FACodec achieves the best performance in the unseen test set, while the system using SimAM-ResNet100 performs best in the seen-speaker test set.

The main contributions of this paper are summarized as follows:

- We propose DSE-vTAD, a novel framework that introduces discriminative speaker embeddings to enhance performance on the vTAD task.
- DSE-vTAD explicitly models the relationship between a pair of speaker embeddings by concatenating them with their Hadamard product.
- Extensive experiments on the vTAD 2025 Challenge show that DSE-vTAD significantly outperforms the baseline systems in unseen and seen-speaker test sets.

2 Task Description

In the vTAD task, a set of timbre attribute descriptors is defined as $\mathcal{V}$. Given a pair of speech utterances $\mathcal{O}_{\mathrm{A}}$ and $\mathcal{O}_{\mathrm{B}}$ from speakers A and B, respectively, the primary goal of vTAD is to determine whether $\mathcal{O}_{\mathrm{B}}$ is stronger than $\mathcal{O}_{\mathrm{A}}$ with respect to a specified timbre descriptor v, where $v \in \mathcal{V}$.

Mathematically, the task can be formulated as a strength comparison hypothesis $\mathcal{H}\left(\langle \mathcal{O}_{\mathrm{A}}, \mathcal{O}_{\mathrm{B}} \rangle, \mathrm{v}\right) \in \{0, 1\}$, where $\mathcal{H} = 1$ indicates that the hypothesis (i.e., $\mathcal{O}_{\mathrm{B}}$ is stronger than $\mathcal{O}_{\mathrm{A}}$ on attribute v) holds true, and $\mathcal{H} = 0$ indicates that the hypothesis is false. The hypothesis is determined by the vTAD algorithm function $\mathcal{F}\left(\langle \mathcal{O}_{\mathrm{A}}, \mathcal{O}_{\mathrm{B}} \rangle | \mathrm{v}; \theta\right)$, with θ representing the set of algorithm parameters.

3 Methods

In this section, we sequentially introduce the proposed DSE-vTAD network, the employed loss functions, and the inference process. The framework of the DSE-vTAD network is illustrated in Fig. 1.

[1] https://github.com/vTAD2025-Challenge/vTAD

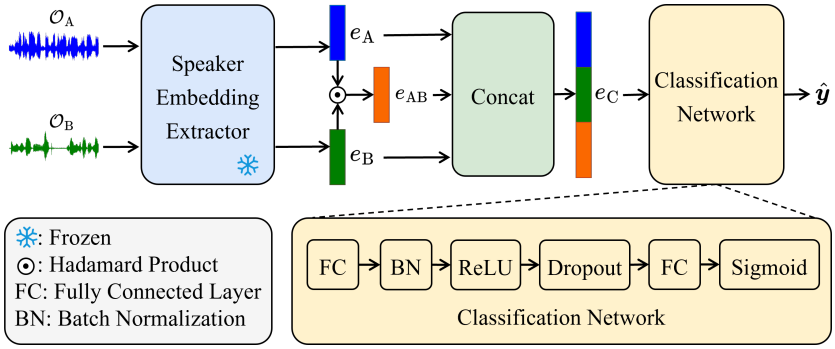

Fig. 1: An overview of the proposed DSE-vTAD network framework.

3.1 DSE-vTAD Network

We define a training sample as $\{\langle \mathcal{O}_A, \mathcal{O}_B \rangle, \boldsymbol{l}\}$, where $\boldsymbol{l}$ is a ground-truth vector label. When the timbre descriptor set $\mathcal{V}$ contains N timbre attributes, $\boldsymbol{l}$ is an N-dimensional vector, where each element $l_n \in \{0, 1, -1\}$ corresponds to the n-th attribute. Specifically, $l_n = 1$ indicates that $\mathcal{O}_B$ is stronger than $\mathcal{O}_A$ in the n-th timbre attribute, while $l_n = 0$ indicates that this comparison does not hold. If $l_n = -1$, the two utterances are not compared under this attribute.

Given a training sample $\{\langle \mathcal{O}_A, \mathcal{O}_B \rangle, \boldsymbol{l}\}$, the DSE-vTAD model first extracts discriminative speaker embeddings $\boldsymbol{e}_A$ and $\boldsymbol{e}_B$ from speech utterances $\mathcal{O}_A$ and $\mathcal{O}_B$, respectively, using a pretrained speaker embedding extractor. This extractor remains frozen throughout the training process. The Hadamard product of $\boldsymbol{e}_A$ and $\boldsymbol{e}_B$ is computed to produce a joint representation $\boldsymbol{e}_{AB}$. These three features—$\boldsymbol{e}_A$, $\boldsymbol{e}_B$, and $\boldsymbol{e}_{AB}$—are concatenated to form a new feature vector $\boldsymbol{e}_C$, which is then passed to a classification network. The classification network adopts the same architecture as used in the vTAD 2025 Challenge: a fully connected layer for dimensionality reduction, followed by a batch normalization layer, a ReLU activation, a dropout layer, a classification fully connected layer, and a final sigmoid function. The output is a prediction vector $\hat{\boldsymbol{y}}$. The predicted value for the n-th timbre attribute descriptor is denoted as $\hat{y}_n$, where $n = 1, 2, \ldots, N$.

3.2 Loss Function

During model training, only samples with labels $l_n \in \{0, 1\}$ are involved in optimizing the model parameters. The loss function is formulated as follows.

$$\mathcal{L} = \mathbb{I}[l_n \in \{0, 1\}] \cdot BCE\left(l_n, \hat{y}_n\right), \tag{1}$$

where $BCE(\cdot)$ denotes the binary cross-entropy function.

3.3 Inference

During the inference stage, given a pair of speech utterances $\langle \mathcal{O}_A, \mathcal{O}_B \rangle$ and the corresponding timbre attribute descriptor v, the utterances $\mathcal{O}_A$ and $\mathcal{O}_B$ are fed

into the DSE-vTAD model to produce a prediction vector $\hat{\boldsymbol{y}}$. The confidence score indicating whether $\mathcal{O}_{\mathrm{B}}$ is stronger than $\mathcal{O}_{\mathrm{A}}$ in the timbre attribute v is obtained from the position in $\hat{\boldsymbol{y}}$ corresponding to v. If the confidence score is greater than or equal to 0.5, the hypothesis is considered true; otherwise, it is false. Furthermore, the confidence scores and the predicted labels can be used to evaluate the model's performance.

4 Experimental Setups

In this section, we first introduce the dataset used in the vTAD 2025 Challenge, followed by a description of the training details of the DSE-vTAD models. Finally, we present the evaluation metrics employed in the experiments.

4.1 Dataset

The VCTK-RVA dataset [5] is used in this challenge. It is built upon the publicly available VCTK corpus [23], with additional annotations that describe the relative intensity differences of voice timbre attributes between same-gender speaker pairs. The dataset includes 18 timbre attributes: bright, thin, coarse, slim, low, pure, rich, magnetic, muddy, hoarse, round, flat, shrill, shriveled, muffled, soft, transparent, and husky. Notably, the attribute husky appears only in male speakers, while shrill appears only in female speakers. As a result, each gender is associated with 17 timbre descriptors. The VCTK-RVA dataset consists of 40,892 speech utterances from 101 speakers and includes 6,038 speaker pair annotations in the form of {Speaker A, Speaker B, voice attribute v}. Each annotation indicates that Speaker B exhibits a stronger intensity than Speaker A in the timbre attribute v. Each speaker pair is annotated with strength comparisons in 1 to 3 timbre attributes.

In the challenge, the VCTK-RVA dataset is divided into two parts for training and evaluation. The training set consists of 29 male and 49 female speakers. For each gender, the speaker pairs annotated in the training set cover all 17 timbre attributes. In total, the training set comprises 136,320 speech utterance pairs. In the test set, all timbre strength comparisons are made within the same gender. For each gender, five timbre attributes are evaluated: the attributes for male speakers are bright, thin, low, magnetic, and pure, while those for female speakers are bright, thin, low, coarse, and slim. Furthermore, based on whether the test speakers appear in the training set, the evaluation defines two tracks: unseen and seen-speaker. For each ordered speaker pair, 20 speech utterances are randomly selected for each speaker, resulting in 400 utterance pairs. The unseen test set contains 91,600 utterance pairs, and the seen-speaker test set contains 94,000. For each timbre attribute in both test sets, the ratio of samples labeled 0 and 1 is 3:1.

4.2 Implementation Details

To extract discriminative speaker embeddings, we adopt four pretrained models: ECAPA-TDNN, FACodec, SimAM-ResNet34, and SimAM-ResNet100. The ECAPA-TDNN[2] and FACodec[3] extractors are consistent with the baseline systems of the vTAD 2025 Challenge. They are trained on the VoxCeleb1&2 [24,25] development sets, containing 7,205 speakers and the Libri-light dataset [26], containing 7,439 speakers, respectively. SimAM-ResNet34 and SimAM-ResNet100 are trained on the large-scale VoxBlink2 dataset [10], which includes 111,284 speakers, using the open-source Wespeaker [27] toolkit[4]. The speaker embeddings extracted by ECAPA-TDNN have a dimensionality of 192, while the others have a dimensionality of 256.

All vTAD models adopt the same training strategy. We use the Adam [28] optimizer with an initial learning rate of 0.01 to update model parameters. A cosine annealing learning rate scheduler is applied to adjust the learning rate for each epoch. The batch size is set to 256, and the training is conducted for a total of 50 epochs. The model obtained from the final epoch is used for evaluation.

4.3 Evaluation Metrics

The challenge adopts Accuracy (ACC) and Equal Error Rate (EER) as performance evaluation metrics. The average ACC and EER, denoted as Avg ACC and Avg EER, respectively, are computed by averaging the ACC and EER scores across all timbre attributes (five attributes for male speakers and five for female speakers, totaling ten attributes). The detailed formulas are as follows:

$$\text{Avg ACC} = \frac{\sum_{i=1}^{N} \text{ACC}_i}{N}, \tag{2}$$

$$\text{Avg EER} = \frac{\sum_{i=1}^{N} \text{EER}_i}{N}. \tag{3}$$

Here, $N = 10$ denotes the total number of timbre attributes. ACC_i and EER_i represent the Accuracy and Equal Error Rate of the i-th timbre attribute, respectively. A higher Avg ACC and a lower Avg EER indicate better system performance.

5 Results and Analysis

In this section, we first compare the performance of the baseline and the proposed DSE-vTAD system. We then report the evaluation results of the DSE-vTAD system using FACodec and SimAM-ResNet100 speaker embedding extractors on each timbre attribute across two test sets. Next, we analyze the effectiveness of incorporating concatenated Hadamard product features. Finally, we conduct an ablation study on the DSE-vTAD system with different dropout rates.

[2] https://github.com/Snowdar/asv-subtools
[3] https://github.com/lifeiteng/naturalspeech3_facodec
[4] https://github.com/wenet-e2e/wespeaker

5.1 Comparison of Baseline and DSE-vTAD Systems Performance

Table 1: Performance comparison overview of different vTAD systems. The best result for each evaluation metric is shown in bold, and the second-best is underlined.

System	Speaker Embedding Extractor	Unseen		Seen-speaker	
		Avg ACC (%)	Avg EER (%)	Avg ACC (%)	Avg EER (%)
Baseline 1	ECAPA-TDNN	71.52	28.33	94.51	5.51
Baseline 2	FACodec	90.76	9.31	93.26	6.45
DSE-vTAD (ours, $p = 0.5$)	ECAPA-TDNN	67.57	32.10	96.02	3.95
	FACodec	<u>90.95</u>	<u>8.92</u>	95.31	4.55
	SimAM-ResNet34	76.08	23.67	96.97	2.86
	SimAM-ResNet100	76.20	23.56	**97.32**	**2.72**
DSE-vTAD (ours, $p = 0.6$)	ECAPA-TDNN	69.25	30.72	95.95	4.13
	FACodec	**91.31**	**8.54**	95.31	4.61
	SimAM-ResNet34	76.39	23.48	97.05	2.98
	SimAM-ResNet100	77.07	22.24	<u>97.15</u>	<u>2.80</u>

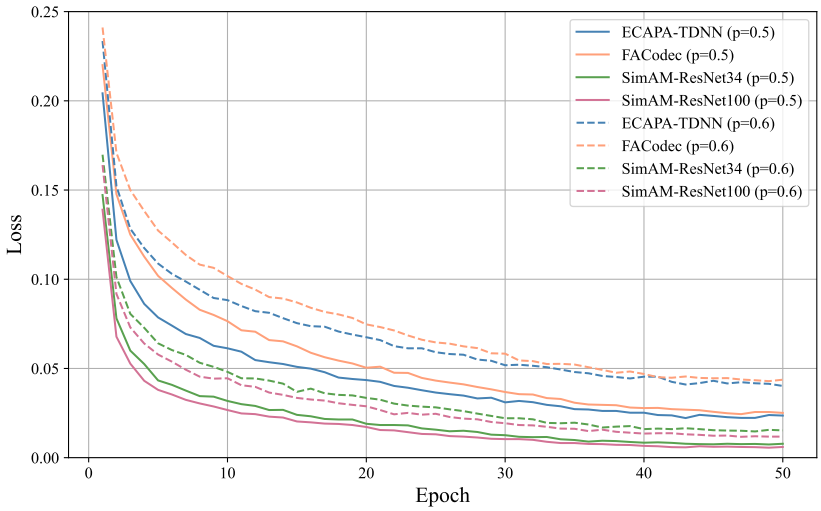

Fig. 2: Training loss curves of different DSE-vTAD systems.

Table 1 compares the Avg ACC and Avg EER of different vTAD systems on the unseen and seen-speaker test sets. The DSE-vTAD system employing the FACodec speaker embedding extractor achieves the best performance on the

unseen test set, while the one using the SimAM-ResNet100 extractor performs best on the seen-speaker test set.

Compared to the baseline system provided by the vTAD 2025 Challenge, the DSE-vTAD system employing the ECAPA-TDNN speaker embedding extractor outperforms the baseline on the seen-speaker test set. With a dropout rate of 0.5, the Avg ACC and Avg EER are relatively improved by 1.60% and 28.31%, respectively. The DSE-vTAD system based on the FACodec extractor surpasses the baseline across all evaluation metrics. For instance, with a dropout rate of 0.6, the Avg ACC and Avg EER on the unseen test set are relatively improved by 0.61% and 8.27%, respectively, while those on the seen-speaker test set are improved by 2.20% and 28.53%, respectively.

Among the DSE-vTAD systems with the same dropout rate, the system utilizing SimAM-ResNet100 as the speaker embedding extractor achieves the best performance on the seen-speaker test set. This can be attributed to its training on large-scale data and larger model capacity. In contrast to the other three speaker embedding extractors trained solely with speaker labels, FACodec employs a speech disentanglement training strategy to reduce the influence of speaker-irrelevant information. As a result, the DSE-vTAD system using FACodec as the extractor achieves the best performance on the unseen test set. These results indicate that speaker embedding extractors with speech disentanglement exhibit better generalization capability in the vTAD task.

Fig. 2 illustrates the variation in training loss over epochs for different DSE-vTAD systems. Since SimAM-ResNet34 and SimAM-ResNet100 are pretrained on larger-scale datasets, the DSE-vTAD systems using them as speaker embedding extractors exhibit lower training losses. To mitigate overfitting, higher dropout rates lead to increased training losses.

Table 2: Evaluation results of the DSE-vTAD systems on the unseen test set. The DSE-vTAD system employing the FACodec speaker embedding extractor is configured with a dropout rate of $p = 0.6$, while SimAM-ResNet100 uses $p = 0.5$.

Speaker Embedding Extractor	**Male**			**Female**		
	Attr.	ACC (%)	EER (%)	Attr.	ACC (%)	EER (%)
FACodec	Bright(明亮)	94.99	5.28	Bright(明亮)	88.94	11.29
	Thin(单薄)	90.42	9.87	Thin(单薄)	89.95	9.85
	Low(低沉)	95.81	4.33	Low(低沉)	85.90	14.82
	Magnetic(磁性)	95.24	3.98	Coarse(粗)	93.16	7.15
	Pure(干净)	85.50	12.72	Slim(细)	93.23	6.06
	Avg	92.39	7.24	Avg	90.24	9.83
SimAM-ResNet100	Bright(明亮)	81.07	20.01	Bright(明亮)	49.27	49.79
	Thin(单薄)	77.41	22.06	Thin(单薄)	48.21	52.76
	Low(低沉)	90.04	7.49	Low(低沉)	55.98	43.78
	Magnetic(磁性)	92.31	6.78	Coarse(粗)	93.04	8.17
	Pure(干净)	83.58	16.39	Slim(细)	91.07	8.38
	Avg	84.88	14.55	Avg	67.51	32.58

Table 3: Evaluation results of the DSE-vTAD systems on the seen-speaker test set. The DSE-vTAD system employing the FACodec speaker embedding extractor is configured with a dropout rate of $p = 0.6$, while SimAM-ResNet100 uses $p = 0.5$.

Speaker Embedding Extractor	**Male**			**Female**		
	Attr.	ACC (%)	EER (%)	Attr.	ACC (%)	EER (%)
FACodec	Bright(明亮)	98.15	1.85	Bright(明亮)	92.74	7.30
	Thin(单薄)	97.88	2.33	Thin(单薄)	94.54	5.27
	Low(低沉)	98.25	1.73	Low(低沉)	99.50	0.47
	Magnetic(磁性)	97.00	3.20	Coarse(粗)	93.11	6.76
	Pure(干净)	83.23	15.90	Slim(细)	98.74	1.26
	Avg	94.90	5.00	Avg	95.73	4.21
SimAM-ResNet100	Bright(明亮)	99.45	0.55	Bright(明亮)	93.24	6.82
	Thin(单薄)	99.42	0.60	Thin(单薄)	96.21	3.76
	Low(低沉)	100.00	0.00	Low(低沉)	98.72	1.28
	Magnetic(磁性)	99.80	0.47	Coarse(粗)	99.58	0.47
	Pure(干净)	86.92	13.13	Slim(细)	99.87	0.14
	Avg	97.12	2.95	Avg	97.52	2.49

5.2 Evaluation Results on Different Timbre Attributes

Tables 2 and 3 present the performance of the DSE-vTAD systems using FACodec and SimAM-ResNet100 as speaker embedding extractors on each timbre attribute in the unseen and seen-speaker test sets, respectively. It can be observed that, on the unseen test set, the strength of timbre attributes is more distinguishable in male speech utterance pairs. In contrast, on the seen-speaker test set, the strength of timbre attributes is more distinguishable in female speech utterance pairs. Furthermore, considering the performance across the unseen and seen-speaker test sets, utterance pairs with the attribute pure for males and bright for females are relatively more difficult to distinguish.

5.3 Effectiveness Analysis of Hadamard Product Features

Table 4 compares the system performance with and without concatenating the Hadamard product features. It can be observed that incorporating the Hadamard product features leads to performance improvements across multiple evaluation metrics. In particular, when the dropout rate is set to 0.5, the DSE-vTAD systems using FACodec and SimAM-ResNet100 as speaker embedding extractors exhibit consistent improvements across all evaluation metrics. Taking SimAM-ResNet100 as an example, the system achieves relative improvements of 0.83% in Avg ACC and 3.80% in Avg EER on the unseen test set, and 0.41% in Avg ACC and 11.11% in Avg EER on the seen-speaker test set.

Table 4: Comparison of vTAD system performance with and without Hadamard product feature concatenation.

Speaker Embedding Extractor	Add e_{AB}	Unseen		Seen-speaker	
		Avg ACC (%)	Avg EER (%)	Avg ACC (%)	Avg EER (%)
$p = 0.5$					
ECAPA-TDNN	✗	**68.53**	**31.97**	95.50	4.20
	✓	67.57	32.10	**96.02**	**3.95**
FACodec	✗	90.33	9.90	95.14	4.92
	✓	**90.95**	**8.92**	**95.31**	**4.55**
SimAM-ResNet34	✗	75.34	25.09	**96.98**	3.17
	✓	**76.08**	**23.67**	96.97	**2.86**
SimAM-ResNet100	✗	75.57	24.49	96.92	3.06
	✓	**76.20**	**23.56**	**97.32**	**2.72**
$p = 0.6$					
ECAPA-TDNN	✗	**70.51**	**29.92**	95.22	4.60
	✓	69.25	30.72	**95.95**	**4.13**
FACodec	✗	90.91	9.39	**95.39**	**4.60**
	✓	**91.31**	**8.54**	95.31	4.61
SimAM-ResNet34	✗	76.11	23.69	96.87	**2.98**
	✓	**76.39**	**23.48**	**97.05**	**2.98**
SimAM-ResNet100	✗	77.06	23.12	**97.11**	2.97
	✓	**77.07**	**22.24**	97.15	**2.80**

5.4 Ablation Study on Dropout Rate

Table 5 compares the performance of the DSE-vTAD system using SimAM-ResNet100 as the embedding extractor under different dropout rates. When the dropout rate is 0.6, the system performs best on the unseen test set, with an Avg ACC of 77.07% and an Avg EER of 22.24%. On the seen-speaker test set, the best performance is observed at a dropout rate of 0.5, with an Avg ACC of 97.32% and an Avg EER of 2.72%. As the dropout rate increases, the system performance improves and degrades. These results suggest that tuning the dropout rate can enhance the robustness and generalization ability of the model.

6 Conclusions

In this paper, we propose a novel framework called DSE-vTAD for the vTAD task. Based on the baseline system of the vTAD 2025 Challenge, DSE-vTAD incorporates discriminative speaker embeddings and concatenates the embeddings of speech utterance pairs along with their Hadamard product features. Experimental results on unseen and seen-speaker test sets demonstrate that DSE-vTAD significantly outperforms the baseline system. In future work, we plan to adopt

Table 5: Comparison of DSE-vTAD system performance across various dropout rates. All systems employ the SimAM-ResNet100 speaker embedding extractor.

Dropout Rate (p)	**Unseen**		**Seen-speaker**	
	Avg ACC (%)	Avg EER (%)	Avg ACC (%)	Avg EER (%)
0.3	75.21	24.33	96.87	3.04
0.4	74.37	25.04	96.69	3.20
0.5	76.20	23.56	**97.32**	**2.72**
0.6	**77.07**	**22.24**	97.15	2.80
0.7	75.97	24.07	96.99	3.19
0.8	76.91	22.80	96.98	3.02

more discriminative speaker embeddings and further optimize the feature fusion strategy to enhance the robustness and generalization ability of the model for the vTAD task.

References

1. He, J., Sheng, Z., Chen, L., Lee, K.A., Ling, Z.H.: Introducing voice timbre attribute detection. arXiv preprint arXiv:2505.09661 (2025)
2. Sheng, Z., He, J., Chen, L., Lee, K.A., Ling, Z.H.: The voice timbre attribute detection 2025 challenge evaluation plan. arXiv preprint arXiv:2505.09382 (2025)
3. Wu, X., Luu, C., Bell, P., Rajan, A.: Explainable attribute-based speaker verification. arXiv preprint arXiv:2405.19796 (2024)
4. Guo, Z., Leng, Y., Wu, Y., Zhao, S., Tan, X.: Prompttts: Controllable text-to-speech with text descriptions. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
5. Sheng, Z.Y., Liu, L.J., Ai, Y., Pan, J., Ling, Z.H.: Voice attribute editing with text prompt. IEEE Transactions on Audio, Speech and Language Processing (2025)
6. Sheng, Z., Du, Z., Lu, H., Zhang, S., Ling, Z.H.: Unispeaker: A unified approach for multimodality-driven speaker generation. arXiv preprint arXiv:2501.06394 (2025)
7. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Interspeech 2017. pp. 999–1003 (2017). https://doi.org/10.21437/Interspeech.2017-620
8. Wang, S., Chen, Z., Lee, K.A., Qian, Y., Li, H.: Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
9. Lin, Y., Qin, X., Zhao, G., Cheng, M., Jiang, N., Wu, H., Li, M.: Voxblink: A large scale speaker verification dataset on camera. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10271–10275. IEEE (2024)
10. Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S., Li, M.: Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. In: Interspeech 2024. pp. 4263–4267 (2024). https://doi.org/10.21437/Interspeech.2024-1490

11. Desplanques, B., Thienpondt, J., Demuynck, K.: Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In: Interspeech 2020. pp. 3830–3834 (2020). https://doi.org/10.21437/Interspeech.2020-2650
12. Yang Zhang and Zhiqiang Lv and Haibin Wu and Shanshan Zhang and Pengfei Hu and Zhiyong Wu and Hung-yi Lee and Helen Meng: MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. In: Interspeech 2022. pp. 306–310 (2022). https://doi.org/{10.21437/Interspeech.2022-563}
13. Qin, X., Li, N., Weng, C., Su, D., Li, M.: Simple attention module based speaker verification with iterative noisy label detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6722–6726. IEEE (2022)
14. Wang, H., Zheng, S., Chen, Y., Cheng, L., Chen, Q.: Cam++: A fast and efficient network for speaker verification using context-aware masking. In: Interspeech 2023. pp. 5301–5305 (2023). https://doi.org/10.21437/Interspeech.2023-1513
15. Liu, S., Song, Z., He, L.: Improving ecapa-tdnn performance with coordinate attention. Journal of Shanghai Jiaotong University (Science) pp. 1–7 (2024)
16. Liu, T., Lee, K.A., Wang, Q., Li, H.: Golden gemini is all you need: Finding the sweet spots for speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2324–2337 (2024)
17. Li, Y., Gan, J., Lin, X., Qiu, Y., Zhan, H., Tian, H.: Ds-tdnn: Dual-stream time-delay neural network with global-aware filter for speaker verification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2814–2827 (2024)
18. Lee, K.A., Wang, Q., Koshinaka, T.: Xi-vector embedding for speaker recognition. IEEE Signal Processing Letters **28**, 1385–1389 (2021)
19. Hong, Q.B., Wu, C.H., Wang, H.M.: Decomposition and reorganization of phonetic information for speaker embedding learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 1745–1757 (2023)
20. Liu, T., Lee, K.A., Wang, Q., Li, H.: Disentangling voice and content with self-supervision for speaker recognition. Advances in Neural Information Processing Systems **36**, 50221–50236 (2023)
21. Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al.: Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 22605–22623 (2024)
22. Cai, D., Li, M.: Leveraging asr pretrained conformers for speaker verification through transfer learning and knowledge distillation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 3532–3545 (2024)
23. Yamagishi, J., Veaux, C., MacDonald, K., et al.: Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). University of Edinburgh. The Centre for Speech Technology Research (CSTR) pp. 271–350 (2019)
24. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: A large-scale speaker identification dataset. In: Interspeech 2017. pp. 2616–2620 (2017). https://doi.org/10.21437/Interspeech.2017-950
25. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Interspeech 2018. pp. 1086–1090 (2018). https://doi.org/10.21437/Interspeech.2018-1929
26. Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al.: Libri-light: A benchmark for asr with limited or no supervision. In: ICASSP 2020-2020 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7669–7673. IEEE (2020)
27. Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., Qian, Y.: Wespeaker: A research and production oriented speaker embedding learning toolkit. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)