The Voice Timbre Attribute Detection 2025 Challenge Evaluation Plan

Zhengyan Sheng[1], Jinghao He[1], Liping Chen[1], Kong Aik Lee[2], Zhen-Hua Ling[1]

[1]NERC-SLIP, University of Science and Technology of China, China
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong
{zysheng, jhhe}@mail.ustc.edu.cn,{lipchen, zhling}@ustc.edu.cn
kongaik.lee@singaporetech.edu.sg

May 13, 2025

1 Challenge Objectives

Voice timbre refers to the unique quality or character of a person's voice that distinguishes it from others as perceived by human hearing. The Voice Timbre Attribute Detection (VtaD) 2025 challenge focuses on explaining the voice timbre attribute in a comparative manner. In this challenge, the human impression of voice timbre is verbalized with a set of sensory descriptors, including bright, coarse, soft, magnetic, and so on. The timbre is explained from the comparison between two voices in their intensity within a specific descriptor dimension. The VtaD 2025 challenge starts in May and culminates in a special proposal at the NCMMSC2025 conference in October 2025 in Zhenjiang, China.

The purpose of this VtaD challenge is to determine whether two voices exhibit obvious intensity differences within a specified descriptor dimension that conceptualizes human impression. The outcomes are expected to uncover the relationship between speech acoustics and human impression of timbre attributes. Furthermore, this task will facilitate explainable speaker recognition [1], serve as an automated voice annotation tool for speaker generation [2, 3, 4], and promote the development of timbre-related speech technologies.

This document describes the challenge task, dataset, and baseline systems that participants can use to build their own VtaD system. Additionally, it provides detailed information on the evaluation metrics and rules, as well as guidelines for registration and submission.

2 Task Definition

As shown in Fig. 1, given a pair of utterances $\mathcal{O}_\mathrm{A}$ and $\mathcal{O}_\mathrm{B}$ from speakers A and B, respectively, and a designated timbre descriptor v, the VtaD evaluates whether the intensity of v in $\mathcal{O}_\mathrm{A}$ is stronger than that in $\mathcal{O}_\mathrm{B}$. Mathematically, the hypothesis about the intensity difference is defined as $\mathcal{H}\left(\mathcal{O}_\mathrm{A}, \mathcal{O}_B, \mathrm{v}\right)$. It means that $\mathcal{O}_\mathrm{B}$ is stronger than $\mathcal{O}_\mathrm{A}$ in the descriptor dimension v. Specifically, $\mathcal{H} \in \{0, 1\}$, where $\mathcal{H} = 1$ indicates that the hypothesis $\mathcal{H}$ is correct, and $\mathcal{H} = 0$ indicates that the hypothesis is incorrect. The hypothesis is determined by the VtaD algorithm function $\mathcal{F}(\mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B}|\mathrm{v}; \theta)$, where θ is the set of algorithm parameters.

3 Evaluation

3.1 Metrics

Evaluations are conducted on speech utterances $\mathcal{O}_\mathrm{A}$ and $\mathcal{O}_\mathrm{B}$, originating from a pair of speakers A and B, respectively. The performance is evaluated in two tasks: verification and recognition. The

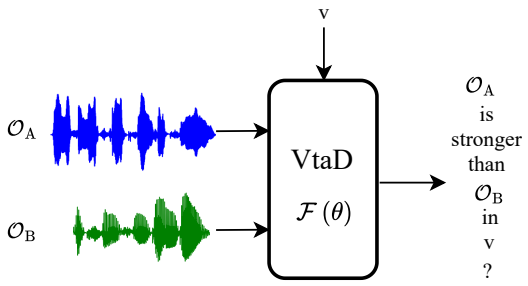

Figure 1: Task definition of VtaD.

hypothesis $\mathcal{H}(\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}) = 1$, where $\mathrm{v} \in \mathcal{V}$, is defined, assuming that $\mathcal{O}_\mathrm{B}$ is stronger than $\mathcal{O}_\mathrm{A}$ in the descriptor dimension v. The system provides the confidence score of $\mathcal{H}$ in the verification evaluation and determines whether $\mathcal{H}$ is correct in the recognition evaluation. The verification results are measured with equal error rate (EER), and the recognition results are measured with accuracy. The lower EERs and higher accuracies indicate better performance.

- *EER*: In the verification evaluation, the target and nontarget trials are composed regarding whether the hypothesis $\mathcal{H}(\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}) = 1$ is true or not. Specifically, the target evaluation samples consist of instances where $\mathcal{H}(\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}) = 1$, while the nontarget samples comprise instances where $\mathcal{H}(\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}) = 0$. Given an evaluation sample $\{\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}\}$, denote the confidence score obtained by the algorithm as $s^{\mathrm{v}}_{\langle \mathrm{A,B} \rangle}$. Higher $s^{\mathrm{v}}_{\langle \mathrm{A,B} \rangle}$ value indicates that $\mathcal{O}_\mathrm{B}$ is more likely to be stronger than $\mathcal{O}_\mathrm{A}$ in the descriptor dimension v. Finally, the EER value is computed on the confidence scores given the ground-truth target and nontarget labels of the evaluations samples.

- *Accuracy (ACC)*: In the recognition evaluation, given the evaluation sample $\{\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}\}$ and the ground-truth label $t \in \{0, 1\}$, a label of 0 indicates that the hypothesis $\mathcal{H}(\langle \mathcal{O}_\mathrm{A}, \mathcal{O}_\mathrm{B} \rangle, \mathrm{v}) = 1$ is false, while a label of 1 indicates that the hypothesis is true. The algorithm predicts whether the hypothesis $\mathcal{H}$ is true or not. Thereby, the accuracy is computed between the prediction and the ground truth t as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{1}$$

  In (1), TP is short for true positives, representing the number of true evaluation samples that are correctly predicted. TN is short for true negatives, representing the number of false evaluation samples that are correctly predicted. FP is short for false positives, denoting the number of false evaluation samples that are incorrectly predicted to be true. FN is short for false negative, denoting the number of true evaluation samples that are incorrectly predicted to be false.

For both EER and ACC, the results obtained by averaging across all evaluated descriptors are used as the indicators of system performance.

3.2 Scenarios

Regarding the speakers applied in the training and evaluation, the performance of timbre attribute intensity detection was conducted in two evaluation scenarios: unseen and seen, as illustrated in Fig. 2. The detailed descriptions are as follows.

- *Unseen*: In the unseen scenario, the speakers used in the evaluation phase are not present in the training phase.

- *Seen*: In this scenario, the speakers employed in the evaluation phase are applied in the training phase, while distinct utterances are utilized for training and evaluation, respectively. Moreover, given a specific speaker, the ordered pairs composed with different speakers are used for training and evaluation.

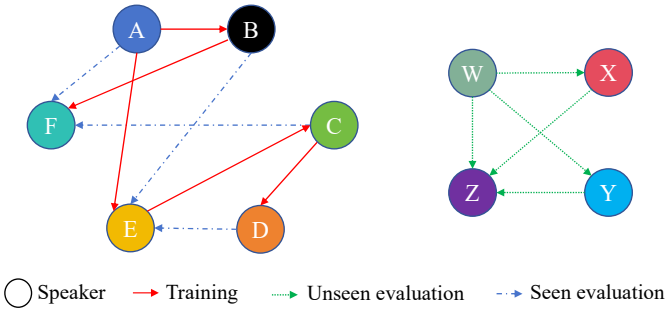

Figure 2: Ordered speaker pair construction in training, unseen and seen evaluations, respectively. The direction of the arrow is from the weaker speaker to the stronger speaker in a specific descriptor in the training annotation and evaluation hypothesis.

4 Data

The VCTK-RVA dataset[5] is employed in our work, wherein the publicly available VCTK database was annotated for timbre intensity. In the dataset, a timbre attribute descriptor set $\mathcal{V}$ is defined, including 18 timbre descriptors, as listed in Table 4. In total, 101 speakers are involved, forming 6,038 annotated ordered speaker pairs {⟨Speaker A, Speaker B⟩, voice attribute v}, indicating that Speaker B is stronger than Speaker A in the specific descriptor dimension v. The number of descriptor dimensions annotated for each ordered speaker pair ranges from 1 to 3. Notably, the speakers are annotated in a gender-dependent manner, utilizing 16 descriptors for both male and female speakers, with each gender assigned one exclusive descriptor.

In this challenge, the VCTK-RVA dataset is partitioned for training and evaluation, respectively. The evaluations are defined in two scenarios regarding whether the test speakers are seen or not in the training. In the seen evaluation scenario, the speakers are included in the training data, while the evaluated speaker pairs are not. For each gender, the training set contains speaker pairs annotated on all 17 voice attributes. In total, 29 male and 49 female speakers are included in the training phase. In the testing phase, five descriptors are selected for each gender. Speaker statistics for the training set, seen test set, and unseen test set are presented in Table 5, Table 6, and Table 7, respectively. In both evaluation datasets, 20 utterances were randomly drawn from each speaker. For a speaker pair, 100 $\mathcal{H} = 1$ and 300 $\mathcal{H} = 0$ evaluation samples are constructed for a descriptor.

5 Baseline Methods

To facilitate the development of customized models by participants, we have publicly released a VtaD framework as illustrated in []. A comprehensive overview of the framework is provided in this section, including model architecture, parameter settings, and training procedures. The baseline performances are also presented. Hopefully, this will serve as a valuable reference for participants in designing and refining their systems, ultimately supporting more effective model optimization and advancing research in the field.

The detailed description for the baseline method can be found in []. Given the utterance pair $\mathcal{O}_\mathrm{A}$ and $\mathcal{O}_\mathrm{B}$, the speaker embedding vectors are extracted with a pre-trained speaker encoder, represented as $\boldsymbol{e}_\mathrm{A}$ and $\boldsymbol{e}_\mathrm{B}$, respectively. Given the speaker embedding vector pair, the model is trained to predict the intensity difference in each timbre descriptor. In this challenge, two speaker encoders are provided in the baseline, including a pre-trained ECAPA-TDNN encoder [6] and the timbre encoder in the FACodec [7] architecture. The details of these two encoders are as follows:

- *ECAPA-TDNN*: The ECAPA-TDNN speaker encoder was trained on the VoxCeleb1 [8] and VoxCeleb2 [9] datasets, utilizing the open-source recipe ASV-Subtools[1].
- *FACodec*: The timbre encoder in the open-source FACodec [10] model[2] was used. It was trained on a 60K-hour Librilight dataset [11].

[1] https://github.com/Snowdar/asv-subtools
[2] https://github.com/lifeiteng/naturalspeech3_facodec

Table 1: Evaluation results of the VtaD model on the unseen speaker test set utilizing the ECAPA-TDNN and FACodec speaker encoders, respectively. The row *Avg* is obtained by averaging the results across all the descriptors for each metric.

Model	Male			Female		
	Attr.	ACC (%)	EER (%)	Attr.	ACC (%)	EER (%)
ECAPA-TDNN	Bright(明亮)	64.46	34.05	Bright(明亮)	47.24	49.83
	Thin(单薄)	72.28	27.63	Thin(单薄)	53.25	49.87
	Low(低沉)	77.71	19.38	Low(低沉)	66.63	35.76
	Magnetic(磁性)	74.31	20.41	Coarse(粗)	91.38	8.42
	Pure(干净)	66.29	34.00	Slim(细)	92.42	7.82
	Avg	71.01	27.10	Avg	70.18	30.34
FACodec	Bright(明亮)	93.24	6.29	Bright(明亮)	88.34	11.70
	Thin(单薄)	95.15	4.93	Thin(单薄)	89.08	10.80
	Low(低沉)	91.52	11.08	Low(低沉)	87.45	13.27
	Magnetic(磁性)	97.88	1.78	Coarse(粗)	91.42	9.12
	Pure(干净)	80.54	19.50	Slim(细)	92.54	6.88
	Avg	91.67	8.72	Avg	89.77	10.35

The evaluation results obtained by the two models in the unseen and seen scenarios are presented in Table 1 and 2, respectively.

6 Evaluation Rules

- Participants are free to develop their own VtaD systems, using components of the baselines or not.
- In addition to the labeled data we provide, participants may use any other data for pre-training, unsupervised learning, and other purposes. Please clearly describe the usage of all data in the final submitted report. Note that, since VCTK is used in our evaluation, only the speakers in the provided training dataset are permitted for algorithm development.
- Participants are allowed to use any-pretrained models and required to describe them clearly in the submission.
- Participants are allowed to make 3 submissions corresponding to different models or training strategies.

7 Registration

Participants/teams are requested to register for the evaluation. Registration should be performed once only for each participating entity using the registration form. Participants will receive a confirmation email within $\sim$ 24 hours after successful registration, otherwise or in case of any questions they should contact the organizers: vtad2025_org@163.com.

8 Submission

We will release the test set before the competition deadline. The format of the test set will consist of pairs of utterances, where participants are required to determine the difference in a specified voice attribute. Additionally, we will provide a reference document outlining the format for submitting feedback results. Participants should submit their results in the specified format before the deadline.

Table 2: Evaluation results of the VtaD model on the seen speaker test set utilizing the ECAPA-TDNN and FACodec speaker encoders, respectively. The row *Avg* is obtained by averaging the results across all the descriptors for each metric.

Model	Male			Female		
	Attr.	ACC (%)	EER (%)	Attr.	ACC (%)	EER (%)
ECAPA-TDNN	Bright(明亮)	94.88	4.78	Bright(明亮)	89.15	9.82
	Thin(单薄)	96.70	3.07	Thin(单薄)	91.37	9.22
	Low(低沉)	99.10	0.93	Low(低沉)	96.35	3.77
	Magnetic(磁性)	96.00	3.03	Coarse(粗)	92.82	7.30
	Pure(干净)	84.50	15.33	Slim(细)	97.39	2.12
	Avg	94.23	5.43	Avg	93.42	6.45
FACodec	Bright(明亮)	95.38	3.97	Bright(明亮)	89.89	9.88
	Thin(单薄)	91.53	8.47	Thin(单薄)	93.09	6.88
	Low(低沉)	96.55	3.77	Low(低沉)	98.64	1.45
	Magnetic(磁性)	96.85	2.90	Coarse(粗)	88.41	12.07
	Pure(干净)	83.03	15.17	Slim(细)	96.81	2.64
	Avg	92.67	6.85	Avg	93.37	6.58

Each participant should also submit a single, detailed system description. All submissions should be made according to the schedule below. Submissions received after the deadline will be marked as late submissions, without exception.

9 Schedule

The result submission deadline is 4th July 2025. All participants are invited to present their work at the special proposal session for VtaD 2025 in NCMMSC2025.

Table 3: Important dates

Challenge Announcement	13th May 2025
Release of the method description and open-source code for the baseline	15th May 2025
Release of training dataset, baseline models	17th May 2025
Release of test set	27th June 2025
Deadline for participants to submit system outputs	4th July 2025
Feedback on the performance evaluation results	11th July 2025
Final paper submission	25th July 2025

References

[1] Xiaoliang Wu et al. "Explainable Attribute-Based Speaker Verification". In: *CoRR* abs/2405.19796 (2024). DOI: 10.48550/ARXIV.2405.19796. arXiv: 2405.19796. URL: https://doi.org/10.48550/arXiv.2405.19796.

[2] Zhifang Guo et al. "Prompttts: Controllable Text-To-Speech With Text Descriptions". In: *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096285. URL: https://doi.org/10.1109/ICASSP49357.2023.10096285.

[3] Zhengyan Sheng et al. "Voice Attribute Editing with Text Prompt". In: *CoRR* abs/2404.08857 (2024). DOI: 10.48550/ARXIV.2404.08857. arXiv: 2404.08857. URL: https://doi.org/10.48550/arXiv.2404.08857.

[4] Zhengyan Sheng et al. “Unispeaker: A Unified Approach for Multimodality-driven Speaker Generation”. In: *CoRR* abs/2501.06394 (2025). DOI: 10.48550/ARXIV.2501.06394. arXiv: 2501.06394. URL: https://doi.org/10.48550/arXiv.2501.06394.

[5] Zheng-Yan Sheng et al. “Voice Attribute Editing With Text Prompt”. In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 1641–1652.

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”. In: *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*. Ed. by Helen Meng, Bo Xu, and Thomas Fang Zheng. ISCA, 2020, pp. 3830–3834. DOI: 10.21437/INTERSPEECH.2020-2650. URL: https://doi.org/10.21437/Interspeech.2020-2650.

[7] Zeqian Ju et al. “NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models”. In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=dVhrnjZJad.

[8] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. “Voxceleb: A large-scale speaker identification dataset”. In: *arXiv preprint arXiv:1706.08612* (2017).

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. “Voxceleb2: Deep speaker recognition”. In: *arXiv preprint arXiv:1806.05622* (2018).

[10] Zeqian Ju et al. “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models”. In: *arXiv preprint arXiv:2403.03100* (2024).

[11] Jacob Kahn 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)*. IEEE. 2020, pp. 7669–7673.

A Tables

Table 4: The descriptor (*Descr.*) set used for describing the timbre. The *Trans.* column gives the corresponding Chinese word. The *Perc.* column presents the percentage (%) of the anontation for each descriptor in the *VCTK-RVA* dataset. The descriptors shrill and husky are exclusively annotated for female and male, respectively.

Descr.	**Trans.**	**Perc.**	**Descr.**	**Trans.**	**Perc.**
Bright	明亮	17.10	Thin	单薄	13.03
Coarse	粗	11.62	Slim	细	11.31
Low	低沉	7.43	Pure	干净	5.48
Rich	厚实	4.71	Magnetic	磁性	3.64
Muddy	浑浊	3.59	Hoarse	沙哑	3.32
Round	圆润	2.48	Flat	平淡	2.15
Shrill(female only)	尖锐	2.08	Shriveled	干瘪	1.74
Muffled	沉闷	1.44	Soft	柔和	0.82
Transparent	通透	0.66	Husky(male only)	干哑	0.59

Table 5: The statistics of the Male and Female speakers in the training set. The number of ordered speaker pairs (*#Pairs*) and the number of speakers (*#Speakers*) are presented for each descriptor (*Descr.*).

Male			Female		
Descr.	**#Pairs**	**#Speakers**	**Descr.**	**#Pair**	**#Speakers**
Bright(明亮)	182	29	Bright(明亮)	428	49
Thin(单薄)	82	29	Coarse(粗)	382	49
Magnetic(磁性)	60	29	Slim(细)	373	49
Low(低沉)	70	26	Thin(单薄)	351	49
Pure(干净)	46	23	Low(低沉)	191	48
Muffled(沉闷)	53	25	Pure(干净)	196	47
Coarse(粗)	64	27	Rich(厚实)	159	47
Muddy(浑浊)	54	27	Hoarse(沙哑)	126	49
Slim(细)	56	27	Muddy(浑浊)	106	44
Shriveled(干瘪)	23	21	Shrill(尖锐)	69	45
Rich(厚实)	24	22	Round(圆润)	35	31
Soft(柔和)	36	24	Flat(平淡)	59	36
Hoarse(沙哑)	26	25	Magnetic(磁性)	44	38
Flat(平淡)	30	23	Shriveled(干瘪)	19	22
Transparent(通透)	10	15	Soft(柔和)	7	14
Husky(干哑)	10	15	Muffled(沉闷)	7	14
Round(圆润)	14	14	Transparent(通透)	2	4

Table 6: The statistics of the Male and Female speakers in the seen test set. The number of ordered speaker pairs (*#Pairs*) and the number of speakers (*#Speakers*) are presented for each descriptor (*Descr.*).

Male			Female		
Descr.	**#Pairs**	**#Speakers**	**Descr.**	**#Pair**	**#Speakers**
Bright(明亮)	20	21	Bright(明亮)	40	39
Thin(单薄)	10	12	Thin(单薄)	35	33
Low(低沉)	10	13	Low(低沉)	20	26
Magnetic(磁性)	10	13	Coarse(粗)	20	26
Pure(干净)	10	13	Slim(细)	20	26

Table 7: The statistics of the Male and Female speakers in the unseen test set. The number of ordered speaker pairs (*#Pairs*) and the number of speakers (*#Speakers*) are presented for each descriptor (*Descr.*).

Male			Female		
Descr.	**#Pairs**	**#Speakers**	**Descr.**	**#Pairs**	**#Speakers**
Bright(明亮)	34	20	Bright(明亮)	35	40
Thin(单薄)	29	10	Thin(单薄)	28	35
Low(低沉)	13	10	Low(低沉)	15	10
Magnetic(磁性)	17	10	Coarse(粗)	26	40
Pure(干净)	6	5	Slim(细)	26	40