QvTAD: Differential Relative Attribute Learning for Voice Timbre Attribute Detection

Zhiyu Wu, Jingyi Fang, Yufei Tang,
Yuanzhong Zheng, Yaoxuan Wang, and Haojun Fei

Qifu Technology, Shanghai, China
`feihaojun-jk@360shuke.com`

Abstract. Voice Timbre Attribute Detection (vTAD) plays a pivotal role in fine-grained timbre modeling for speech generation tasks. However, it remains challenging due to the inherently subjective nature of timbre descriptors and the severe label imbalance in existing datasets. In this work, we present QvTAD, a novel pairwise comparison framework based on differential attention, designed to enhance the modeling of perceptual timbre attributes. To address the label imbalance in the VCTK-RVA dataset, we introduce a graph-based data augmentation strategy that constructs a directed acyclic graph and employs disjoint-set union techniques to automatically mine unobserved utterance pairs with valid attribute comparisons. Our framework leverages speaker embeddings from a pretrained FACodec, and incorporates a Relative Timbre Shift-Aware Differential Attention module. This module explicitly models attribute-specific contrasts between paired utterances via differential denoising and contrast amplification mechanisms. Experimental results on the VCTK-RVA benchmark demonstrate that QvTAD achieves substantial improvements across multiple timbre descriptors, with particularly notable gains in cross-speaker generalization scenarios.

Keywords: Voice Timbre Attribute Detection · Speech Perception · Differential Relative Attribute Learning.

1 Introduction

Timbre attribute modeling is critical for advancing speech generative technologies. As a core element of conversation, a speaker's timbre conveys both identity-related characteristics and unique vocal style—shaped by vocal tract anatomy, age, emotion, and environment. These variations shape auditory perception and generate diverse expressive impressions. For generative tasks like speech synthesis, developing highly expressive, finely controllable multi-speaker systems requires multi-dimensional objective modeling of synthesized timbre attributes. However, modeling timbre variations across styles remains challenging due to the subjective and abstract nature of timbre descriptions.

Some speech generation algorithms, such as Text-to-Speech (TTS) and Voice Conversion (VC), have seen studies conduct in-depth exploration into the description and extraction of speakers' timbre features. Seed-VC [8] enhances timbre representation using a diffusion-based Transformer architecture, while introducing training-time perturbations to mitigate timbre leakage, thereby achieving more accurate voice conversion. In the TTS domain, NaturalSpeech3 [7] proposed a codec-based speech representation that decomposes waveform signals into prosody, linguistic content, acoustic details, and timbre of speaker. This disentangled representation improves the naturalness and fidelity of reference speaker synthesis. Meanwhile, controllable TTS via instruction-based conditioning has gained increasing attention, enabling control over emotion, speed, and energy through natural language prompts, as demonstrated by [5][16][17]. Despite these advances, fine-grained modeling of timbre remains highly challenging, and tasks such as precise timbre analysis and controllable timbre generation are still in their early stages of exploration [1][11].

To enable timbre-aware editing and comparison, several recent studies have attempted to capture and manipulate differences in speaker timbre attributes. For example, VoxEditor [10] introduces a manually annotated dataset, VCTK-RVA, which encodes relative attribute differences between speakers to support text-guided voice attribute editing. This dataset provides a controllable and reproducible benchmark for timbre modeling. Building on this foundation, the newly proposed task of vTAD, as detailed in [12] and [4], focuses on pairwise comparisons of timbre attribute strengths based on the

Husky
Muffled
Soft
Coarse
Flat
Hoarse
Rich
Low
Slim

Shriveled
Transparent
Round
Bright
Shrill
Magnetic
Muddy
Pure
Thin

(b) Mel-spectrogram of two speech clips from two different speakers.

(a) Annotated Timbre Attributes in the VCTK-RVA Dataset.

(c) Analysis of different ratio of F0, High frequency, Jitter and Shimmer.
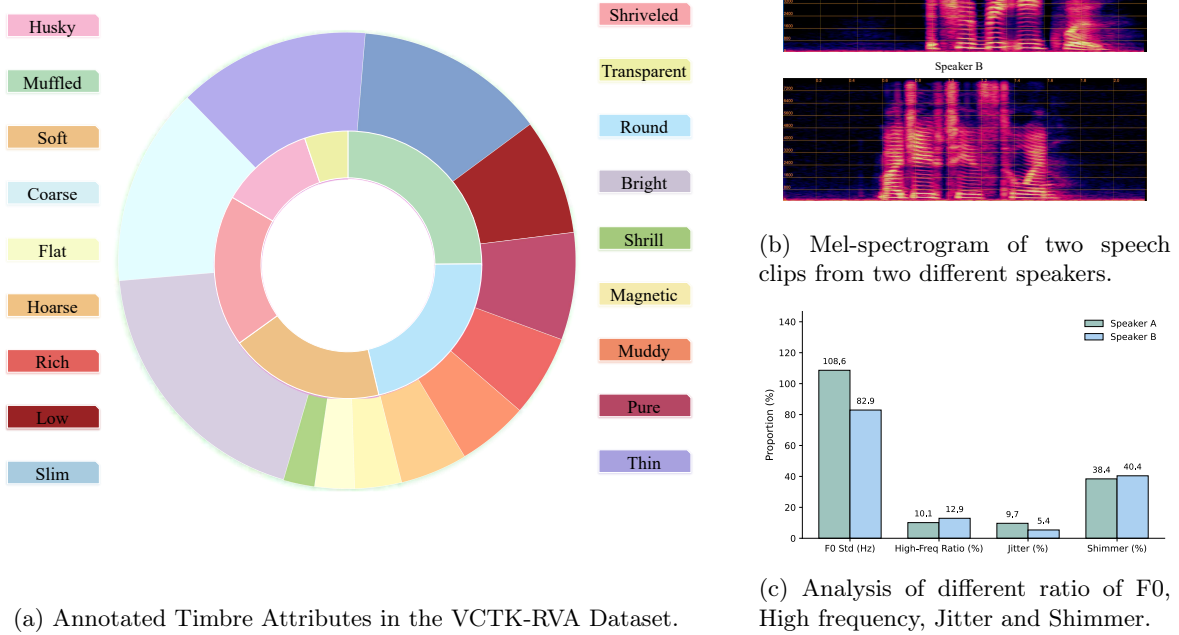
Fig. 1: Introduction to Voice Attribute and case analysis of *Mellow* attribute

VCTK-RVA dataset. Details regarding the dataset and task setup are introduced in Section 4.1.1 and Section 3.1.1, respectively.

In this paper, we propose a novel differential attention-based pairwise comparison framework for more precise speaker timbre modeling. First, based on extensive analysis of the VCTK-RVA dataset, we design an automated graph-based data augmentation method to address imbalanced timbre attribute label distribution. Second, we leverage a pretrained speech codec model for speaker embedding extraction and introduce the $RTSA^2$(Relative Timbre Shift-Aware Differential Attention) module—a differential denoising mechanism—to learn attribute-specific contrasts between utterances from different speakers. Finally, we train the model on VCTK-RVA to predict which utterance exhibits stronger presence of a given timbre attribute, with comparative experiments and ablation studies demonstrating improved accuracy and generalization to out-of-set scenarios.

Our main contributions are:

- We propose an automated graph-based data augmentation method, which constructs a Directed Acyclic Graph (DAG) via Disjoint-Set Union over VCTK-RVA to mitigate timbre attribute label imbalance.
- We propose $RTSA^2$ structure, which models timbre embedding differences and incorporates a contrastive predictor to improve attribute strength classification accuracy both within and outside the dataset.
- Experimental validation that our QvTAD method enhances prediction accuracy across multiple timbre descriptors, with notable gains in unseen-speaker settings.

2   Related Work

2.1   Cross-Domain Relative Attribute Modeling

The task of "given two samples and an attribute, determine which sample is stronger in that attribute" is prevalent across various domains and is widely known as relative attribute learning. This formulation avoids the subjectivity and inconsistency of scalar annotations by instead supervising models through

comparisons between pairs of samples. It has been applied successfully in Computer Vision, Natural Language Processing, and medical imaging, etc.

In Computer Vision, Parikh [9] first introduced the concept of relative attributes using image pairs to define attribute strength relationships, and trained a ranking SVM accordingly. Souri proposed Deep Relative Attributes [18], leveraging CNNs for feature extraction and optimizing a ranking loss in an end-to-end framework. Zhang [20] introduced DACRL, incorporating attribute centers in the embedding space to improve ranking interpretability. Relative modeling of style or emotion intensity has gained traction in Natural Language Processing. Jhamtani [6] trained a style transfer model using pairs of Shakespearean and modern sentences to model stylistic strength.

These methods share a unified formulation: sample pair + attribute + strength comparison, emphasizing perceptual-space modeling through relational supervision. Our task of predicting timbre strength in speech extends this paradigm to the auditory domain.

2.2 Relative Perception and Timbre Attribute Detection

While relative attribute learning is mature in Computer Vision and NLP, it remains underexplored in speech. Wester's MOSPC framework trains ranking models from subjective speech naturalness comparisons, outperforming traditional MOS in consistency and reliability [14]. NR-SQA introduced *pairwise and triplet ranking losses* for no-reference speech quality assessment, directly modeling ordinal perceptual differences to improve generalization to synthetic and real speech [13].

Despite these advances, speech attribute modeling faces two central challenges: (1) Abstraction: Timbre perception involves nonlinear interactions of frequency structure, harmonic content, and energy distribution, making it difficult to judge from Mel spectrograms. (2) Subjectivity: Timbre annotation requires expert knowledge and lacks consistency, limiting scalability.

Recent studies by VoxEditor [10] introduced VCTK-RVA—the first timbre relative modeling dataset-featuring pairwise comparisons of 34 attributes (e.g., magnetic, bright, hoarse) annotated by human raters. Its attribute composition (Figure 1a) makes it one of the few public resources for speech-based relative attribute learning.

However, timbre's inherent abstraction and subjectivity lead to listener discrepancies. Conventional signal-based analyses often mismodel timbre: for instance, "soft" voices (typically with smooth mid/low-frequency energy) may contradict low-level features. As shown in Figure 1b–c, Speaker B was rated "softer" by speech AI experts despite conflicting acoustic features (higher high-frequency ratio, shimmer)—highlighting limitations of low-level descriptors.

Prior work on pathological voice assessment, which models extensive acoustic parameters [2], suffers from complexity, inefficiency, and sensitivity to dataset bias, lacking robustness at scale.

3 Method

This paper focuses on the task of modeling the relative strength of timbre-related attributes in speech, aiming to determine which of two given speech samples exhibits a stronger presence of a target timbre attribute. To this end, we propose QvTAD, a novel model that integrates a differential attention mechanism with a perceptual difference amplification strategy for relative attribute prediction. Our approach follows the paradigm of "sample pair + attribute word + relative strength judgment" and is composed of a pairwise comparision oriented data augmentation method with Disjoint-Set Searching and a main prediction model with two core components: (1) a Differential Attention Prior-Guided Relative Attribute Amplification Mechanism, which enhances the perception of fine-grained attribute differences between sample pairs, and (2) a multi-label classifier conditioned on the target attribute. As illustrated in 2, the overall architecture comprises three main modules: a frozen FACodec Encoder for timbre representation extraction, a differential attention module for contrastive modeling, and a final probabilistic prediction network with sigmoid activation.

3.1 Data Augmentation with Disjoint-Set Searching

We analyzed an existing speech attribute comparison dataset (see Section 4.1 for VCTK details), which includes multiple attribute dimensions (e.g., brightness, roughness) defined by speaker pairwise comparisons.
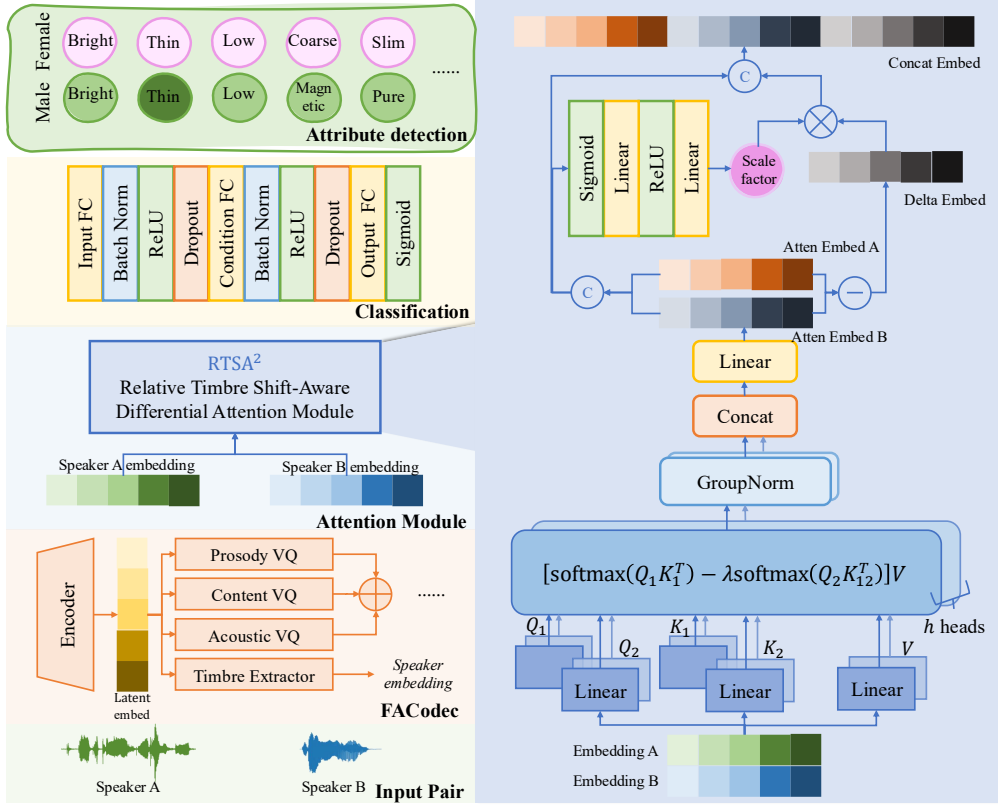
Fig. 2: Overview of the proposed QvTAD framework.

Statistical analysis reveals a pronounced long-tail distribution: head attributes (e.g., *Brightness_F*, *Rough_F*) account for over 45% of samples, while tail attributes (e.g., *Transparent_F*, *Gentle_F*) are severely underrepresented (some <1%). This imbalance may bias training, hindering learning of rare attributes.

Notably, we observed a key structural property: for any speaker pair and attribute, all comparisons share the same label direction (no contradictions). This suggests the dataset is built on speaker-level rather than clip-level rankings, providing valuable priors for downstream tasks like graph-based modeling.

To address the issue of imbalanced data distribution across timbre attributes and insufficient coverage of training pairs, we propose a Disjoint-Set Union (DSU)-based data augmentation strategy tailored for pairwise comparison tasks. Inspired by structured modeling, our method abstracts each speaker and their associated timbre attribute strength as nodes and edges within a directed graph, enabling the construction of an attribute-aware graph structure that systematically discovers potential, yet unobserved, comparable pairs to enrich the training data.

Specifically, each annotated pair in the training set is represented as a directed edge between two nodes of the form ⟨Speaker, Attribute⟩. For example, if utterance $\mathcal{O}_B$ is annotated as stronger than $\mathcal{O}_A$ with respect to attribute v_i, we add a directed edge from node (A, v_i) to (B, v_i), indicating the relative strength relation. All annotated pairs form a directed graph under each attribute dimension. Under the assumption of transitivity in attribute comparison, this graph is expected to form a locally ordered DAG.

Building on this structure, we introduce an attribute-specific DSU-based augmentation strategy. For each attribute $v_i \in \mathcal{V}$, we maintain an independent disjoint-set structure to track transitive connectivity among nodes under that attribute. We then iterate over all unannotated node pairs (u, v) and query their connectivity in the DSU: if a transitive path exists between u and v but the pair is not present in the original training set, the pair is treated as a pseudo-labeled candidate pair. To mitigate the risk of noisy or inconsistent annotations inherent to graph construction, we apply a minimum path length

threshold and a multi-path voting mechanism to ensure the reliability and quality of the augmented pairs.

3.2 Relative Timbre Shift-Aware Differential Attention

Input Modeling and Task Formulation We consider as input a pair of speech embeddings, denoted as $\mathbf{e}_A, \mathbf{e}_B \in \mathbb{R}^d$, where $d = 256$ is the embedding dimensionality. The timbre attribute of interest is specified by an attribute index vector $\mathbf{l} \in \{0,1\}^K$, where $K = 34$ denotes the number of predefined timbre attributes. The objective is to predict whether $\mathbf{e}_B$ exhibits a stronger presence of the specified attribute compared to $\mathbf{e}_A$. The model outputs a scalar probability $p \in [0,1]$, representing the confidence that "sample B is stronger than sample A" for the given attribute.

Following NaturalSpeech3 [7], we employ a pretrained FACodec Encoder to extract 1024-dimensional speech embeddings for each audio sample. The encoder is frozen during training to maintain stable representations.

Differential Attention Prior-Guided Relative Attribute Amplification Given the nature of relative attribute modeling—centered around contrastive perception—we introduce a differential attention mechanism to enhance representation differences between samples. This module employs multi-head differential denoising attention [19] to suppress shared noise in the attention maps between the query-key pairs, thereby strengthening discriminative contrast features.

We begin by concatenating the two input embeddings into a sequence:

$$\mathbf{E} = [\mathbf{e}_A; \ \mathbf{e}_B] \in \mathbb{R}^{2 \times d} \tag{1}$$

This sequence is fed into the multi-head differential attention module to generate enhanced contrastive representations: $\mathbf{e}_A^{\mathrm{att}}, \ \mathbf{e}_B^{\mathrm{att}} \in \mathbb{R}^d$. Inspired by differential amplifiers, the module computes attention differences between two sets of query-key pairs to suppress common-mode signals and amplify significant contrastive cues. Specifically, each attention head computes:

$$[Q_1; \ Q_2] = \mathbf{E}W_Q, \quad [K_1; \ K_2] = \mathbf{E}W_K \tag{2}$$

The attention difference is then computed as:

$$\mathbf{A}\mathrm{diff} = \mathrm{softmax}\left(\frac{Q_1 K_1^\top}{\sqrt{d}}\right) - \lambda \cdot \mathrm{softmax}\left(\frac{Q_2 K_2^\top}{\sqrt{d}}\right) \tag{3}$$

where $\lambda \in (0,1)$ is a learnable scaling factor (initialized as λ_{init}) to balance the contribution of the two attention maps. The resulting attention map $\mathbf{A}_{\mathrm{diff}}$ is applied to the value vectors V, yielding the attention-enhanced contrastive representations $\mathbf{e}_A^{\mathrm{att}}, \ \mathbf{e}_B^{\mathrm{att}}$.

Since relative attribute modeling always operates on a pair of samples, the input sequence length remains fixed at 2. As such, our model focuses on pairwise contrast rather than long-range contextual dependency, and thus positional encoding is unnecessary. Furthermore, we empirically observed that Rotary Positional Embeddings (RoPE) did not improve performance in this setting and introduced unnecessary computation. We therefore exclude RoPE from the attention head to simplify the model and improve training efficiency.

To further enhance the model's sensitivity to subtle timbre differences, we compute a difference vector, which is non-linearly modulated and normalized to generate an amplified contrastive representation:

$$\Delta = \mathbf{e}_B^{\mathrm{att}} - \mathbf{e}_A^{\mathrm{att}}; \hat{\Delta} = \tanh(\Delta) \cdot \|\Delta\|_2 \cdot \gamma \tag{4}$$

where $\gamma \in [0,2]$ is a learnable amplification factor that scales the perceptual contrast based on the difficulty of the sample pair. We empirically found this bounded range leads to stable training and better performance. The amplification factor γ is generated by a contrastive prediction module:

$$\gamma = \sigma\left(f_{\mathrm{scale}}([\mathbf{e}_A^{\mathrm{att}}; \ \mathbf{e}_B^{\mathrm{att}}])\right) \tag{5}$$

where $f_{\mathrm{scale}}(\cdot)$ is a two-layer feedforward network with non-linear activations, and $\sigma(\cdot)$ denotes the sigmoid function. The output is rescaled to the range $[0,2]$ to control the intensity of contrast amplification.

Finally, we concatenate three sources of information—$\mathbf{e}_A^{\mathrm{att}}$, $\mathbf{e}_B^{\mathrm{att}}$, and $\hat{\Delta}$—and feed them into a feed-forward prediction network composed of two fully connected layers, batch normalization, and dropout for regularization. The output is a probability vector:

$$\mathbf{o} = \sigma(W\mathrm{out}\mathbf{z}) \in \mathbb{R}^K \tag{6}$$

where the k-th element o_k represents the predicted probability that sample B is stronger than sample A with respect to attribute k. During training, we apply supervision only on the target attribute dimension where $l_k = 1$, using a binary cross-entropy (BCE) loss.

4   Experiments and Evaluation

4.1   Dataset

For the vTAD task, we utilize the VCTK-RVA dataset, which is derived from the publicly available CSTR VCTK Corpus [15] through additional annotation. The original VCTK corpus consists of speech recordings from 110 English speakers with diverse accents. Each speaker reads approximately 400 sentences, selected from a newspaper article, the Rainbow Passage, and an elicitation paragraph used in the Speech Accent Archive.

VCTK-RVA provides relative strength annotations for voice attributes between speech segments from different speakers of the same gender. The training set involves 101 speakers. For each ordered pair of speakers, 1 to 3 timbre descriptors are annotated, resulting in a total of 6,038 labeled pairs in the form of Speaker A, Speaker B, voice attribute v, indicating that Speaker B exhibits a stronger presence of attribute v compared to Speaker A.

After data augmentation, the training set contains 166,409 samples. For validation and testing, we use the original splits provided with the dataset. The validation set consists of 1,085 samples. The test set is divided into two subsets: seen (94,000 samples) and unseen (91,600 samples), representing speakers included and excluded from the training set, respectively. All audio samples are downsampled to 16 kHz.

4.2   Training Setup

For fair evaluation on unseen speakers, we adopt the pretrained codec model in NaturalSpeech3, FACodec as the timbre encoder which produces 256-dimensional representations. FaCodec is trained on the Librilight corpus, which ensures no speaker overlap with the VCTK test set.

Model training is performed on 8 NVIDIA A800 GPUs via the Accelerate framework [3], with a batch size of 8,192 samples per GPU. The model is trained for 20 epochs until convergence. We employ the Adam optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, with a learning rate of 5×10^{-5}.

For baseline evaluation, we follow the official configuration provided by the vTAD Challenge, using both the dataset and training/inference protocols as described in the challenge paper and official codebase [1].

4.3   Evaluation and Ablation Analysis

In this section, the QvTAD model is compared with the reproduced results of the official FAcodec baseline model of the vTAD 2025 challenge [12], with the original data from the paper listed above. Following the evaluation framework proposed in [10] [12], accuracy (ACC) and equal error rate (EER) are adopted as the evaluation metrics, and a systematic comparison is conducted on the performance of different models across two gender categories (Male, Female) and two scenarios (Seen, Unseen). Among them, QvTAD-AST corresponds to the model data of the competition scheme, which is included solely to ensure the fairness of the comparison; QvTAD-RTSA² represents the proposed model in this paper. The experimental results demonstrate that in the Seen scenario, QvTAD-RTSA² achieves the second-highest accuracy for both genders, with its performance surpassed only by QvTAD-AST and FACodec (Reproduced). In the Unseen scenario, the proposed model attains best performance across

Table 1: Performance comparison across different speaker groups on VCTK-RVA testset.

Method	Seen (M/F)			Unseen (M/F)		
	ACC (%) ↑	EER (%) ↓	Avg ACC	ACC (%) ↑	EER (%) ↓	Avg ACC
ECAPA-TDNN (Reported)	94.23/93.42	5.43/6.45	93.83	71.01/70.18	27.10/30.28	70.60
FACodec (Reported)	92.67/93.37	6.85/6.58	93.02	91.67/89.77	8.72/10.38	90.72
FACodec (Reproduced)	**87.49**/84.57	**12.17**/14.92	86.03	81.55/70.42	18.04/29.66	<u>75.99</u>
QvTAD-AST	84.95/**88.79**	15.55/**12.03**	**86.87**	74.00/<u>76.44</u>	26.59/<u>25.30</u>	75.22
QvTAD-RTSA^2	<u>86.45</u>/<u>85.32</u>	<u>13.82</u>/14.93	85.89	**91.90**/**82.07**	**7.89**/**18.52**	**86.99**

all evaluation metrics, fully validating the superior capability of the proposed method in learning speaker-irrelevant differences.

To evaluate the effectiveness of each proposed component in our model, we conduct ablation experiments by progressively removing key modules and comparing performance on both seen and unseen speaker groups. Table 2 summarizes the results.

We first evaluate the impact of our DSU Searching-based data augmentation. Removing it leads to significant performance degradation, particularly for seen speakers (-2.12% ACC) and to a lesser extent for unseen speakers (-1.44% ACC). This confirms that augmentation enhances the model's robustness to intra- and inter-speaker variations, promoting better generalization. Next, we assess our RTSA² module, designed to accentuate perceptual differences between embedding pairs while suppressing shared timbral features. Removing this module causes a slight drop in accuracy for unseen speakers (-0.71%) despite a marginal gain for seen speakers (+0.10%), suggesting RTSA² is critical for generalizing to unseen speaker distributions. This is likely attributed to its contrastive learning mechanism, which emphasizes relative timbre shifts over absolute representations.

Overall, the full QvTAD-RTSA^2 model achieves the best performance across both seen and unseen conditions, validating the effectiveness of both components. The results highlight the importance of contrast-aware attention and data diversity in relative timbre attribute modeling.

Table 2: Ablation study of module effectiveness between seen and unseen speaker groups.

Method	Seen ACC (%)	Unseen ACC (%)
QvTAD-RTSA^2	85.89	86.99
w/o Data Augmentation	83.77	85.55
w/o RTSA^2 Module	85.99	86.28

5   Conclusion

We introduce QvTAD, a novel differential attention-based framework for modeling voice timbre attributes through pairwise comparisons. By integrating a graph-based data augmentation method and a shift-aware differential attention mechanism, QvTAD effectively addresses the challenges of label imbalance and perceptual subjectivity in timbre modeling. Experiments on the VCTK-RVA dataset demonstrate that QvTAD achieves superior performance across a range of timbre descriptors, especially in cross-speaker evaluation. These results highlight the potential of comparative comparison of voice timbre attributes and fine-grained acoustic modeling for advancing perceptual attribute understanding in speech generation. Future work may extend this framework to multi-lingual timbre analysis or real-time controllable synthesis.

References

1. Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Ni, C., Shi, X., An, K., Yang, G., Li, Y., Chen, Y., Gao, Z., Chen, Q., Gu, Y., Chen, M., Chen, Y., Zhang, S., Wang, W., Ye,

---

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

J.: Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training (2025), https://arxiv.org/abs/2505.17589

2. Farrús, M., Hernando, J., Ejarque, P.: Jitter and shimmer measurements for speaker recognition. In: Interspeech 2007. pp. 778–781 (2007). https://doi.org/10.21437/Interspeech.2007-147

3. Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., Sun, M., Bossan, B.: Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate (2022)

4. He, J., Sheng, Z., Chen, L., Lee, K.A., Ling, Z.H.: Introducing voice timbre attribute detection (2025), https://arxiv.org/abs/2505.09661

5. Huang, R., Hu, R., Wang, Y., Wang, Z., Cheng, X., Jiang, Z., Ye, Z., Yang, D., Liu, L., Gao, P., Zhao, Z.: Instructspeech: following speech editing instructions via large language models. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)

6. Jhamtani, H., Gangal, V., Hovy, E., Nyberg, E.: Shakespearizing modern language using copy-enriched sequence to sequence models. In: Brooke, J., Solorio, T., Koppel, M. (eds.) Proceedings of the Workshop on Stylistic Variation. pp. 10–19. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4902, https://aclanthology.org/W17-4902/

7. Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S.: Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)

8. Liu, S.: Zero-shot voice conversion with diffusion transformers (2024), https://arxiv.org/abs/2411.09943

9. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision. pp. 503–510 (2011). https://doi.org/10.1109/ICCV.2011.6126281

10. 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 **33**, 1641–1652 (2025). https://doi.org/10.1109/TASLPRO.2025.3557193

11. Sheng, Z., Du, Z., Lu, H., Zhang, S., Ling, Z.H.: Unispeaker: A unified approach for multimodality-driven speaker generation (2025), https://arxiv.org/abs/2501.06394

12. Sheng, Z., He, J., Chen, L., Lee, K.A., Ling, Z.H.: The voice timbre attribute detection 2025 challenge evaluation plan (2025), https://arxiv.org/abs/2505.09382

13. Ta, B.T., Le, M.T., Do, V.H., Thanh Binh, H.T.: Enhancing no-reference speech quality assessment with pairwise, triplet ranking losses, and asr pretraining. In: Interspeech 2024. pp. 2700–2704 (2024). https://doi.org/10.21437/Interspeech.2024-2527

14. Wang, K., Zhao, Y., Dong, Q., Ko, T., Wang, M.: MOSPC: MOS prediction based on pairwise comparison. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1547–1556. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-short.132, https://aclanthology.org/2023.acl-short.132/

15. Yamagishi, J., Veaux, C., MacDonald, K.: CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). https://doi.org/10.7488/ds/2645 (2019), university of Edinburgh. The Centre for Speech Technology Research (CSTR)

16. Yang, D., Liu, S., Huang, R., Weng, C., Meng, H.: Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2913–2925 (2024). https://doi.org/10.1109/TASLP.2024.3402088

17. Yang, G., Yang, C., Chen, Q., Ma, Z., Chen, W., Wang, W., Wang, T., Yang, Y., Niu, Z., Liu, W., Yu, F., Du, Z., Gao, Z., Zhang, S., Chen, X.: Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting (2025), https://arxiv.org/abs/2504.12867

18. Yang, X., Zhang, T., Xu, C., Yan, S., Hossain, M.S., Ghoneim, A.: Deep relative attributes. IEEE Transactions on Multimedia **18**(9), 1832–1842 (2016). https://doi.org/10.1109/TMM.2016.2582379

19. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=OvoCm1gGhN

20. Zhang, Z., Li, Y., Zhang, Z.: Relative attribute learning with deep attentive cross-image representation. In: Zhu, J., Takeuchi, I. (eds.) Proceedings of The 10th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 95, pp. 879–892. PMLR (14–16 Nov 2018), https://proceedings.mlr.press/v95/zhang18d.html