CUHK-EE-DSP&STL Systems for the vTAD Challenge at NCMMSC 2025

Aemon Yat Fei Chiu[1], Jingyu Li[2], Tan Lee[1], Yusheng Tian[1], and Guangyan Zhang[2]

[1] Dept of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[2] Independent Researcher
{aemon.yf.chiu, lijingyu0125}@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk, {ystian0617,gyzhang}@link.cuhk.edu.hk

Abstract. This paper presents the Voice Timbre Attribute Detection (vTAD) systems developed by the Digital Signal Processing & Speech Technology Laboratory (DSP&STL) at The Chinese University of Hong Kong (CUHK) for the 20th National Conference on Human-Computer Speech Communication (NCMMSC 2025) vTAD Challenge. The proposed systems leverage WavLM-Large embeddings with attentive statistical pooling to extract robust speaker representations, followed by two variants of Diff-Net–Feed-Forward Neural Network (FFN) and Squeeze-and-Excitation-enhanced Residual FFN (SE-ResFFN)–to compare timbre attribute intensities between utterance pairs. Experimental results demonstrate that the WavLM-Large+FFN system generalises better to unseen speakers, achieving 77.96% accuracy and 21.79% EER, while the WavLM-Large+SE-ResFFN model excels in the 'Seen' setting with 94.42% accuracy and 5.49% EER. These findings highlight a trade-off between model complexity and generalization, and underscore the importance of architectural choices in fine-grained speaker modelling. Our analysis also reveals the impact of speaker identity, annotation subjectivity, and data imbalance on system performance, pointing to future directions for improving robustness and fairness in timbre attribute detection.

Keywords: Voice timbre attribute detection · Speaker verification · WavLM · Squeeze-and-excitation · Speech representation learning.

1 Introduction

The Voice Timbre Attribute Detection (vTAD) Challenge [1, 2], held as part of the 20th National Conference on Human-Computer Speech Communication (NCMMSC 2025), focuses on identifying perceptual differences in voice timbre attributes between speakers by comparing pairs of utterances. Timbre is characterised using a set of sensory descriptors inspired by various modalities, including auditory (e.g., hoarse, rich), visual (e.g., bright, dark), tactile (e.g., soft, hard), and physical (e.g., magnetic, transparent) perceptions.

Our proposed systems leverage WavLM-Large [3], a large-scale speech self-supervised learning (SSL) representation model, to extract robust voice features. These features are further refined using attentive statistical pooling (ASTP) [4] before being passed into a comparison network, referred to as Diff-Net. We explore two architectural variants of Diff-Net for modelling attribute intensity differences: a standard Feed-Forward Neural Network (FFN) [5], and a Deep Residual FFN enhanced with Squeeze-and-Excitation blocks (SE-ResFFN) [5–8].

The design motivation is to harness the rich, hierarchical representations encoded by large-scale speech self-supervised learning (SSL) models, while effectively utilising deep speaker verification (SV) architectures as downstream classifiers for fine-grained timbre attribute comparison.

2 Datasets

The core component of our system is WavLM-Large, a self-supervised model pre-trained on a massive 94,000-hour corpus comprising 10,000 hours from GigaSpeech [9], 24,000 hours from VoxPopuli [10], and 60,000 hours from LibriLight [11].

For training and evaluation of the downstream Diff-Net modules (i.e., FFN and SE-ResFFN), we use the VCTK-RVA dataset [12]. This dataset provides annotations for 17 timbre descriptors for both male and female speakers, resulting in a total of 34 distinct timbre attributes. Each descriptor captures a perceptual quality of the voice, and comparisons are made between pairs of utterances to determine relative intensity.

The VCTK-RVA dataset comprises over 6,000 annotated speaker pairs, with each pair labelled for one to three timbre descriptors. These descriptors characterise perceptual qualities of the voice, such as bright, thin, coarse, magnetic, shrill, and husky. Since the annotations are derived from human judgments, they inherently reflect subjective interpretations of vocal timbre. This subjectivity introduces variability in the data, which poses challenges for model consistency and generalisation across unseen speakers and utterances.

3 System Description

3.1 Audio Pre-Processing

To enhance the accuracy of timbre attribute extraction, we apply a silence removal procedure that eliminates non-informative segments from the audio signal. Silence is defined as regions where the signal energy falls below 40 dB, and is detected using a sliding window of 25 milliseconds (ms) with a hop size of 10 ms. The process targets leading and trailing silence while preserving internal pauses, thereby retaining the natural rhythm and structure of speech. To prevent excessive trimming, a safeguard is implemented: if the remaining waveform is shorter than 100 ms, the silence removal step is bypassed to ensure the preservation of meaningful acoustic content.

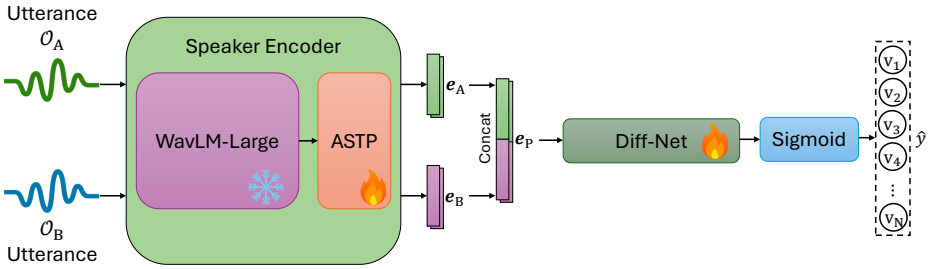

Fig. 1. The overall design concept of the systems.

3.2 System Overview

Figure 1 illustrates the overall architecture of our proposed systems. The design closely follows the baseline framework introduced in the vTAD Challenge [1, 2], which consists of a speaker encoder followed by a comparison network.

Given a pair of input utterances, denoted as $\mathcal{O}_A$ and $\mathcal{O}_B$, the speaker encoder extracts corresponding embeddings $\mathbf{e}_A$ and $\mathbf{e}_B$. These embeddings are concatenated to form a joint representation $\mathbf{e}_P$, which is then passed into the Diff-Net module.

The Diff-Net produces an output vector of dimension N, where each element corresponds to a specific timbre attribute. A sigmoid activation is applied to this output to generate the prediction vector $\hat{y}$. The n-th element of $\hat{y}$ ($n = 1, 2, ..., 34$) represents the predicted likelihood that utterance $\mathcal{O}_B$ exhibits a stronger intensity than $\mathcal{O}_A$ for the n-th timbre descriptor.

3.3 Speaker Encoder

Figure 2 illustrates the speaker encoder architecture, which integrates the WavLM-Large model [3] with an ASTP module [4]. The WavLM-Large model consists of a convolutional neural network (CNN) encoder layer [13] followed by 24 stacked Transformer blocks [14]. Each pre-processed utterance is fed into the model in its entirety, without cropping.

From the CNN encoder and each of the 24 Transformer blocks, we extract 1024-dimensional frame-level intermediate representations, resulting in 25 such frame-level representations per utterance. These representations are first averaged across frames to produce 25 layer-wise embeddings, which are then aggregated by the ASTP module.

Originally designed for frame-level aggregation in SV tasks, the ASTP module is repurposed in our system to perform layer-wise aggregation. This adaptation allows the model to exploit the rich hierarchical features encoded across all layers of the WavLM-Large model. Specifically, eight attention heads are employed to compute weighted statistics (mean and standard deviation) across layers, yielding a 2048-dimensional embedding for each utterance.

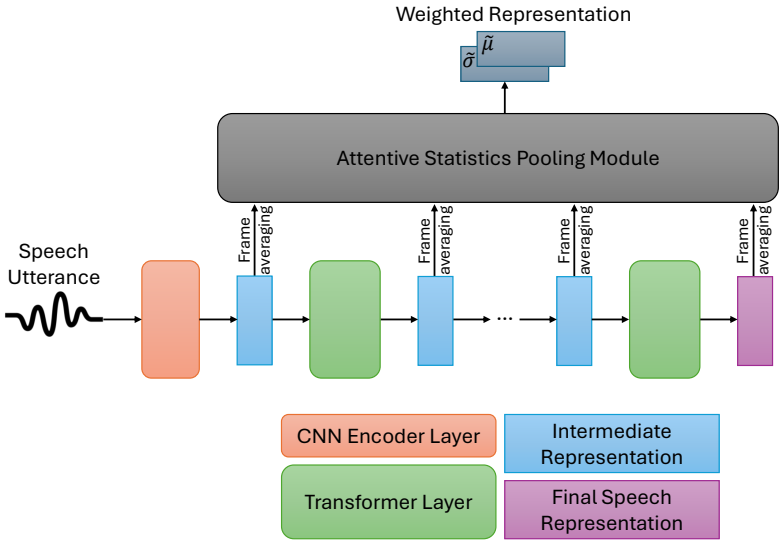

Fig. 2. The WavLM-Large module with the adoption of ASTP for voice feature extraction.

The embeddings from the two input utterances are concatenated to form a 4096-dimensional feature vector, which serves as input to the Diff-Net. To mitigate over-fitting, two dropout layers with a dropout rate of 0.1 are incorporated within the ASTP module.

3.4 Diff-Net

Figure 4 presents the architectures of the two Diff-Net variants employed in our systems.

The first variant is an enhanced version of the traditional Feed-Forward Neural Network (FFN) used in prior work [1, 2]. It consists of four sequential FFN blocks with hidden dimensions of [512, 256, 128, 64].. Each block comprises a fully connected (FC) layer, followed by batch normalisation (BN), a ReLU activation function, and a dropout layer with a dropout rate of 0.3. After the final FFN block, an additional FC layer is applied to produce predictions for all 34 timbre attributes.

The second variant, SE-ResFFN, is inspired by the squeeze-and-excitation ResNet (SE-ResNet) architecture, which has shown strong performance in SV tasks [6, 8]. As illustrated in Figure 5, this model incorporates four SE-ResNet blocks with hidden dimensions of [1024, 1024, 512, 256]. Following these blocks, a BN layer is applied, and the resulting features are passed through two FC layers with hidden dimensions [192, 64], interleaved with a ReLU activation. A final FC layer then outputs the predictions for the 34 timbre attributes.

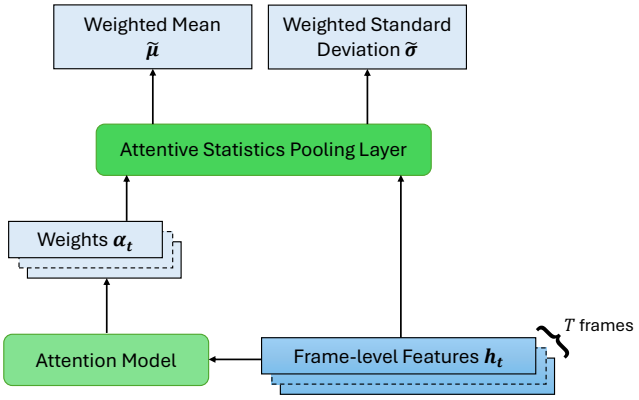

Fig. 3. The original ASTP mechanism.

Both architectures conclude with a sigmoid activation function, which transforms the raw outputs into probability scores representing the predicted likelihood that the second utterance exhibits a stronger intensity for each timbre attribute.

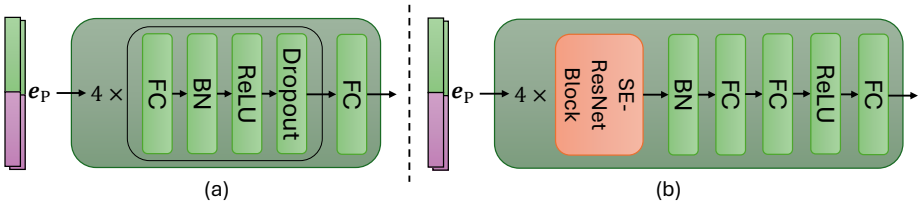

Fig. 4. The architecture of the Diff-Net structure based on (a) FFN and (b) SE-ResFFN.

4 Excremental Settings

Model training was conducted over 10 epochs with a batch size of 16, using the AdamW optimiser [15] with a learning rate of $1e-4$ and a weight decay of 0.01. A cosine annealing scheduler [16] was employed to adjust the learning rate dynamically throughout training. The loss function used was binary cross-entropy with sample-wise reduction, and model validation was performed after each epoch. To ensure reproducibility, all experiments were initialised with a fixed random seed (42).

The proposed systems were evaluated under two distinct scenarios: 'Unseen' and 'Seen'. In the 'Unseen' scenario, the speakers in the test set do not appear in

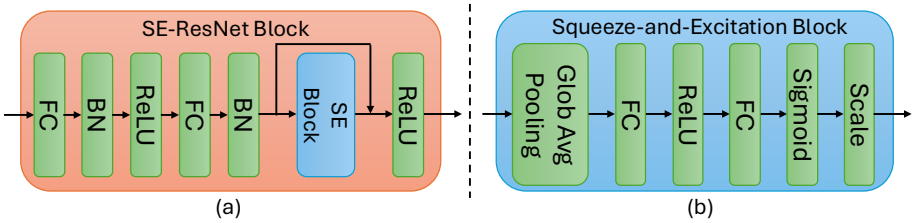

Fig. 5. (a) An SE-ResNet block. (b) A squeeze-and-excitation block inside an SE-ResNet block.

the training set, thereby assessing the model's generalisation capability. In the 'Seen' scenario, the same set of speakers is used for both training and evaluation, but with non-overlapping utterances. Furthermore, for each target speaker, pairings with other individuals are uniquely assigned to either the training or evaluation set to prevent overlap.

Performance is measured using two standard metrics: accuracy (Acc) and equal error rate (EER). Accuracy reflects the proportion of correct predictions, while EER indicates the point at which false acceptance and false rejection rates are equal. Higher accuracy and lower EER values correspond to better system performance.

Table 1. The results on the test set

Test Set	Model	Acc (%)	EER (%)
Unseen	WavLM-Large +**FFN**	**77.96**	**21.79**
	WavLM-Large +**SE-ResFFN**	74.90	25.17
Seen	WavLM-Large +**FFN**	90.77	9.08
	WavLM-Large +**SE-ResFFN**	**94.42**	**5.49**

5 Results and Analysis

Table 1 summarises the overall performance of our systems under two evaluation scenarios: 'Seen' and 'Unseen'. As expected, both models perform substantially better on the 'Seen' test set, where speakers are present in the training data, than on the 'Unseen' set, which evaluates generalisation to novel speakers. For instance, the WavLM-Large+SE-ResFFN system achieves 94.42% accuracy and

5.49% EER on 'Seen', but drops to 74.90% accuracy and 25.17% EER on 'Unseen'. This gap highlights the difficulty of disentangling timbre attributes from speaker identity, suggesting potential over-fitting to speaker-specific patterns.

Table 2. Evaluation results of our proposed systems on the 'Unseen' test set. The row **Avg** is obtained by averaging the results across all the descriptors for each metric.

Model	Male			Female		
	Attribute	Acc (%)	EER (%)	Attribute	Acc (%)	EER (%)
WavLM-Large +**FFN**	Bright (明亮)	69.12	31.82	Bright (明亮)	58.56	41.43
	Thin (單薄)	73.71	24.67	Thin (單薄)	55.57	44.01
	Low (低沉)	83.35	15.97	Low (低沉)	71.45	28.13
	Magnetic (磁性)	94.94	5.35	Coarse (粗)	89.85	10.06
	Pure (乾淨)	88.79	10.61	Slim (細)	88.79	5.81
	Avg	81.98	17.69	Avg	73.94	25.89
WavLM-Large +**SE-ResNet**	Bright (明亮)	65.49	32.52	Bright (明亮)	49.33	48.27
	Thin (單薄)	74.03	26.47	Thin (單薄)	48.65	52.62
	Low (低沉)	90.67	10.05	Low (低沉)	70.02	31.71
	Magnetic (磁性)	80.31	19.80	Coarse (粗)	88.65	12.24
	Pure (乾淨)	88.38	11.72	Slim (細)	93.40	6.33
	Avg	79.78	20.11	Avg	70.01	30.23

Interestingly, the WavLM-Large+FFN model outperforms WavLM-Large+SE-ResFFN in the 'Unseen' setting, while WavLM-Large+SE-ResFFN excels in the 'Seen' setting. This contrast may stem from the architectural complexity and inductive biases of the models. The SE-ResFFN-based Diff-Net, with its deeper layers and squeeze-and-excitation mechanisms, is better equipped to capture fine-grained speaker-specific patterns, which benefits performance when the test speakers are seen during training. However, this same specialisation may hinder generalisation to novel speakers, as the model may over-fit to specific traits rather than learning robust, speaker-invariant representations. In contrast, the simpler FFN-based Diff-Net architecture may generalise better due to its lower model complexity and reduced reliance on dynamic feature weighting, which may help it capture more speaker-invariant patterns and thus perform more robustly in the 'Unseen' scenario. This observation suggests a trade-off between model expressiveness and generalisation, and highlights the importance of tailoring architectural choices to the target deployment context.

To further probe model robustness, we conducted small-scale experiments using speaker-disjoint training splits and observed considerable variation in results depending on the speaker composition. This, along with the inherent subjectivity in manual timbre labelling, suggests that model performance is sensitive to both speaker identity and annotation consistency. Additionally, the dataset [12] exhibits imbalance in descriptor frequency and gender representation–some

Table 3. Evaluation results of our proposed systems on the 'Seen' test set.

Model	Male			Female		
	Attribute	Acc (%)	EER (%)	Attribute	Acc (%)	EER (%)
WavLM-Large +**FFN**	Bright (明亮)	93.38	7.03	Bright (明亮)	85.06	14.97
	Thin (單薄)	92.85	6.53	Thin (單薄)	90.14	9.34
	Low (低沉)	96.05	3.37	Low (低沉)	93.26	6.63
	Magnetic (磁性)	95.25	4.80	Coarse (粗)	86.06	13.80
	Pure (乾淨)	79.40	20.33	Slim (細)	96.25	3.95
	Avg	91.39	8.41	Avg	90.15	9.74
WavLM-Large +**SE-ResNet**	Bright (明亮)	96.66	2.83	Bright (明亮)	85.61	12.96
	Thin (單薄)	97.42	3.00	Thin (單薄)	89.21	10.79
	Low (低沉)	98.10	1.30	Low (低沉)	97.49	3.32
	Magnetic (磁性)	99.62	0.40	Coarse (粗)	90.60	9.66
	Pure (乾淨)	91.88	8.47	Slim (細)	97.55	2.20
	Avg	96.74	3.20	Avg	92.09	7.78

attributes are under-represented, and female speakers dominate the data, yet male speakers often yield better results. These factors collectively highlight the need for more balanced data and refined annotation practices to improve generalisation and fairness.

Detailed results for the 'Unseen' and 'Seen' test sets, broken down by gender and descriptor, are listed in Table 2 and Table 3, respectively.

6 Conclusion

This paper presents the CUHK-EE-DSP&STL systems submitted to the vTAD Challenge at NCMMSC 2025, designed to detect perceptual differences in voice timbre attributes through pairwise utterance comparison. Our approach integrates WavLM-Large embeddings with attentive statistical pooling for robust speaker representation, followed by two variants of Diff-Net, i.e., FFN and SE-ResFFN, for attribute intensity comparison. The systems achieved second place in the 'Unseen' track and fourth place in the 'Seen' track, demonstrating competitive performance across both generalisation and speaker-specific scenarios.

The results highlight the rich representational capacity of speech SSL models like WavLM, especially when paired with carefully designed downstream architectures. Notably, the FFN model showed stronger generalisation to novel speakers, while SE-ResFFN excelled in capturing fine-grained patterns among known speakers, suggesting a trade-off between model complexity and robustness. Our analysis also underscores the challenges posed by speaker variability, annotation subjectivity, and data imbalance, pointing to key areas for future improvement. These findings pave the way for further research in fine-grained speaker modelling and voice attribute disentanglement using self-supervised speech representations.

References

1. Sheng, Z., He, J., Chen, L., Lee, K. A., Ling, Z.-H.: The Voice Timbre Attribute Detection 2025 Challenge Evaluation Plan. arXiv:2505.09382 (2025). https://doi.org/10.48550/arXiv.2505.09382
2. He, J., Sheng, Z., Chen, L., Lee, K. A., Ling, Z.-H.: Introducing Voice Timbre Attribute Detection. arXiv:2505.09661 (2025). https://doi.org/10.48550/arXiv.2505.09661
3. Chen, S., et al.: WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022). https://doi.org/10.1109/JSTSP.2022.3188113
4. Okabe, K., Koshinaka, T., Shinoda, K.: Attentive Statistics Pooling for Deep Speaker Embedding. In: Proceedings of Interspeech 2018, pp. 2252–2256. (2018). https://doi.org/10.21437/Interspeech.2018-993
5. Bebis, G., Georgiopoulos, M.: Feed-Forward Neural Networks. IEEE Potentials **13**(4), 27–31 (1994). https://doi.org/10.1109/45.329294
6. Zeinali, H., Wang, S., Silnova, A., Matějka, P., Plchot, O.: WavLM: BUT System Description to VoxCeleb Speaker Recognition Challenge 2019. arXiv:1910.12592 (2019). https://doi.org/10.48550/arXiv.1910.12592
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. (2016). https://doi.org/10.1109/CVPR.2016.90
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. (2018). https://doi.org/10.1109/CVPR.2018.00745
9. Chen, G., et al.: GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In: Proceedings of Interspeech 2021, pp. 3670–3674. (2021). https://doi.org/10.21437/Interspeech.2021-1965
10. Wang, C., et al.: VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 993–1003. (2021). https://doi.org/10.18653/v1/2021.acl-long.80
11. Kahn, J., et al.: WavLM: Libri-Light: A Benchmark for ASR with Limited or No Supervision. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7669–7673. (2020). https://doi.org/10.1109/ICASSP40776.2020.9052942
12. Vaswani, A., et al.: Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'17). 6000–6010 (2017). https://doi.org/10.5555/3295222.3295349
13. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM **60**(6), 84–90 (2012). https://doi.org/10.1145/3065386
14. 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
15. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: Proceedings of the International Conference on Learning Representations (ICLR) 2019. (2019).
16. Loshchilov, I., Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts. In: Proceedings of the International Conference on Learning Representations (ICLR) 2017. (2017).