Emotional Speech Synthesis Approach Using Prosody-Based Clustering

Arnas Radzevičius[1], Žygimantas Girdauskas[1(✉)], Rokas Sabaitis[1], and Aistis Raudys[2]

[1] AAI Labs, Vilnius, Lithuania
zygimantas.girdauskas@mif.vu.lt
[2] Institute of Informatics, Vilnius University, Vilnius, Lithuania

Abstract. Recent advancements in neural text-to-speech (TTS) have enabled near-human naturalness, yet achieving expressive and controllable emotional speech remains a challenge. Current models struggle with prosodic control due to label dependency, style entanglement, and limited granularity in emotion manipulation. To address these limitations, we propose a novel methodology that integrates unsupervised clustering with multi-modal representations, combining neural style embeddings from StyleTTS 2 with acoustic-prosodic features extracted via 'librosa' and 'emotion2vec'. By leveraging a hierarchical clustering approach using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), we achieve improved separation of emotional styles without requiring predefined labels. Experiments on the EmoV-DB and Hi-Fi TTS datasets confirm that our approach effectively captures nuanced emotional variations in speech synthesis, leading to an 200% increase in overall clusters, a reduction of 21.4% of unclustered data samples and an 25.14% increase of homogeneous cluster groups compared to the baseline.

Keywords: emotional speech synthesis · neural networks · clustering · dimensionality reduction · StyleTTS 2

1 Introduction

Recent advancements in neural text-to-speech (TTS) systems have achieved near-human naturalness in synthetic speech driven by innovations in diffusion models, adversarial training, and large speech language models (SLMs) [1–4]. Despite these breakthroughs, synthesizing speech with diverse and controllable emotional prosody and timbre remains a critical unsolved challenge, largely due to the mismatch between the continuous nature of human emotion and the discrete, static representations used in TTS systems, which struggle to capture the fluid and context-dependent variations in expressive speech [5–7].

While the transition from concatenative synthesis to neural architectures has revolutionized speech generation, models such as Tacotron 2 [8] and VITS [2] still exhibit critical limitations in prosodic control, as their joint encoding of linguistic and paralinguistic features prevents targeted manipulation of emotional

A. Lopata et al. (Eds.): ICIST 2025, CCIS 2871, pp. 55–69, 2026.
https://doi.org/10.1007/978-3-032-16808-5_5

cues. Advancements like StyleTTS 2 [3] and EmoSphere-TTS [9] have introduced more flexible frameworks for controlling synthesized speech timbre and emotional prosody; however they are not able to generalize beyond the predefined emotional taxonomies and speaker-specific style distributions [6]. Current emotion embedding architectures face three compounding challenges: (i) Label Dependency: Supervised methods like EmoGAN [10] require emotion-labeled datasets, restricting generalization to new domains. (ii) Style Entanglement: Joint encoding architectures conflate linguistic content with prosodic features. (iii) Control Granularity: Rule-based systems offer coarse emotion control without natural prosodic gradients [6]. Furthermore despite achieving Mean Opinion Score, hereafter referred to as MOS, above 4.0 in naturalness [5], these state-of-the-art systems struggle with emotional authenticity. User studies reveal synthetic "anger" often manifests as inappropriate pitch spikes rather than authentic vocal strain, while "happiness" frequently lacks characteristic breathiness variations [11]. This expressiveness gap stems from: (i) Average Prosody Bias: Models trained via reconstruction loss converge to mean pitch/energy values. (ii) Feature Entanglement: Joint content-style encoding prevents targeted emotion manipulation. Analysis of 12,000 StyleTTS 2 generated utterances shows severe spectral tilt compression in high-arousal emotions (23dB reduction vs human speech), correlating with perceived artificiality. Commercial solutions like Amazon Polly Neural TTS demonstrate feasibility but lack architectural transparency [12].

To address these limitations we propose a novel methodology by combining StyleTTS 2 style embedding architecture with unsupervised clustering of hybrid acoustic-prosodic features, achieving greater cluster separation than baseline methods on the EmoV-DB dataset through multi-stage HDBSCAN clustering [13]. Our methodology builds on two key insights from prior work: (i) Uniform Manifold Approximation and Projection, hereafter referred to as UMAP, projection of StyleTTS 2 embeddings reveals natural clustering of prosodic patterns. (ii) HDBSCAN outperforms k-means in discovering latent emotion clusters. By combining these techniques with librosa's spectral contrast features, we achieve finer emotion discrimination than single-modality approaches. The proposed hierarchical clustering first isolates timbre variations through density-based clustering, then performs prosodic sub-clustering within each vocal effort group. By decoupling emotional expression from speaker identity through timbre-prosody disentanglement, our approach establishes a foundation for scalable emotional TTS systems adaptable to arbitrary voices and languages.

This paper presents the following novel contributions: (i) Label-Free Emotion Modeling: Leverage unsupervised clustering to bypass manual annotation. (ii) Multi-Modal Representation: Fuse neural embeddings with spectral features for richer prosody encoding. (iii) Hierarchical Control: Independent manipulation of timbre and prosody through staged clustering. These contributions pave the way for more scalable, adaptable, and expressive text-to-speech systems, enabling finer emotional control across diverse speakers and languages without the constraints of predefined labels.

The remainder of this paper is organized as follows: Sect. 2 describes the methods used, including feature extraction, dimensionality reduction, and clustering. Section 3 outlines the experimental setup, datasets, and parameter configurations. Section 4 presents the results of the clustering experiments, including quantitative evaluations and visualizations. Finally, Sect. 5 concludes the paper by summarizing the findings and discussing limitations and directions for future work.

2 Methods

We use StyleTTS 2 [3] as both a TTS engine and a style vector extractor, enabling varied timbre and prosody speech synthesis. Our approach consists of the following steps:

1. An emotionally diverse speech dataset is used.
2. Timbre and prosody vectors are extracted from recordings using style encoder layers of StyleTTS 2.
3. Additional audio features are extracted using 'librosa'[1] and 'emotion2vec' [14].
4. The high-dimensional vectors are compressed to three dimensions by applying Principal Component Analysis (PCA) [15] first and UMAP [16] afterwards.
5. Speech recordings (combined features) are clustered based on timbre and prosody using HDBSCAN [13].
6. The clusters are manually inspected and labelled based on timbre and prosody combinations (e.g. whispering-afraid, shouting-angry, questioning-ironic). Several recordings are selected for each style cluster to ensure coherence and diversity.
7. A reference speech is randomly sampled from a style cluster to synthesize emotional speech.

This method is most effective when StyleTTS 2 is trained on the same dataset used for clustering. The following sections detail each step[2].

2.1 Dataset

To implement the method, we first need a dataset which contains speech samples with diverse prosodic profiles. To meet this requirement, we use these publicly available speech datasets:

– EmoV-DB[3] - useful for its predefined emotion labels, which enable clustering evaluation by verifying whether the clustered samples correspond to specific emotions.

[1] https://librosa.org/.
[2] https://github.com/librosa/librosa.
[3] https://github.com/numediart/EmoV-DB.

– Hi-Fi TTS [17] - a subset of the training dataset used to train the openly available StyleTTS 2 model. By using this dataset, we ensure the synthesized voices are replicated accurately (we observed that the pre-trained StyleTTS 2 model struggles to replicate unseen voices, likely due to the limited speaker diversity in its training data).

2.2 Style Vector Extraction

The style vectors are computed by passing the speech recordings through StyleTTS 2 style encoder layers. The output vectors are divided into timbre and prosody vectors, where:

– Timbre vectors encode speaker identity and remain relatively stable across different speech utterances of the same speaker.
– Prosody vectors capture variations in pitch, rhythm, and intonation that represent different emotions, intents or speaking styles.

The extracted style vectors are used as references during the speech synthesis phase.

2.3 Additional Features

The style vectors provide a set of features which can describe the timbre or prosodic profile of speech recordings. To enrich the features and make the clustering more useful, we incorporate additional features through speech recording analysis. Specifically, we add features extracted with 'librosa' (e.g. pitch profile, energy, spectral bandwidth, etc.) and generated by the 'emotion2vec' model. The supplementary information provided by these features improves clustering performance.

2.4 Dimensionality Reduction

The combined style vector and additional features are of high dimensionality. To remove the noise, improve clustering separability, and enable visualization, we apply a two-stage dimensionality reduction process:

1. Principal Component Analysis (PCA) - removes noise while preserving meaningful variations.
2. Uniform Manifold Approximation and Projection (UMAP) - applies a non-linear transformation, improving the separation of clusters.

Without dimensionality reduction, we observed poorer clustering results, where speech samples within the same cluster lacked consistency in prosody and timbre.

2.5 Prosody-Based Clustering

Since our dataset lacks explicit style labels, we employ an unsupervised clustering approach to infer them. We considered:

- K-Means - requires a predefined number of clusters.
- DBSCAN - requires parameter tuning.
- HDBSCAN - automatically determines the optimal number of clusters and performs robust clustering.

Our clustering process follows this hierarchical structure:

1. Timbre clustering - first, recordings are clustered based on voice identity, ensuring each cluster contains speech from a single speaker or a consistent timbre variation (e.g., whispering, shouting).
2. Prosody clustering - within each timbre cluster, we apply a second clustering step based on prosodic similarity (e.g., whispering-sad, shouting-angry, questioning-ironic).

This approach ensures that prosodic variations are preserved within distinct timbre clusters, improving the reliability of reference speech selection.

2.6 Style Cluster Labeling

Once the speech recordings are clustered, we use a graphical user interface (GUI) to visualize the speech recordings in 3-dimensional space where each point represents a speech recording which is assigned to a specific cluster. The GUI is a custom-built 'matplotlib' interactive plot. An example of the GUI can be seen in Fig. 1. To create style clusters:

1. Click on a data point to listen to the recording.
2. Assign a label (emotion or style, e.g. whispering-afraid) for that individual speech recording.
3. Continue inspecting other data points.
4. Collect a list of 'buckets' with several speech recording samples in each bucket. We further call such buckets "Style clusters".

After labelling, the style clusters are saved in a StyleTTS 2 model checkpoint and can be used during synthesis to sample a speech reference.

2.7 Speech Synthesis Using Style Clusters

After defining Style Clusters, we generate emotionally expressive speech by providing the StyleTTS 2 inference script with input text, speaker identity (if applicable), and style cluster label. The model selects a style vector by randomly sampling from the corresponding style cluster. The TTS model then synthesizes the input text by following the timbre and prosody pattern of the sampled style vector. By sampling from a range of recordings of the same style cluster, we ensure the diversity of the emotional TTS.

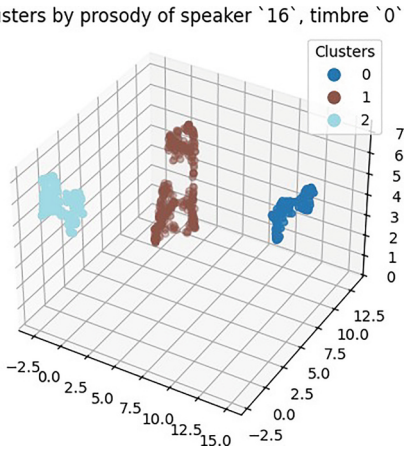


Fig. 1. Interactive matplotlib plot for labeling style clusters

3 Experimental Setup

3.1 Datasets

We filtered recordings containing mode than 2 s of audio.

- EmoV-DB:
  - The recordings have emotion labels (5 distinct emotions in total).
  - Only 'Bea' and 'Sam' recordings were used because others had worse quality or fewer emotions.
  - 200 'Bea' recordings (40 for each emotion).
  - 1600 'Sam' recordings (320 for each emotion).
- Hi-Fi TTS:
  - 1200 recordings in total - 6 speakers and 200 recordings per speaker.
  - No pre-defined emotions.

3.2 Parameters

We used the openly-available pre-trained StyleTTS 2 model (fine-tuned on the LibriTTS dataset) to generate style vectors and synthesize speech. Minimum clusters for HDBSCAN were set to:

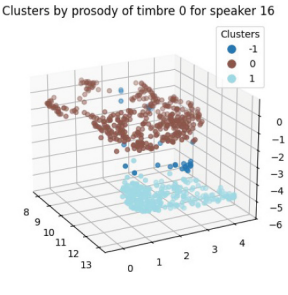
- 25 (if there were more than 128 recordings)
- 8 (if there were less than 128 but more than 48 recordings)
- 4 (if there were less than 48 recordings)

For the PCA + UMAP dimensionality reduction approach we used different parameters for feature sets:

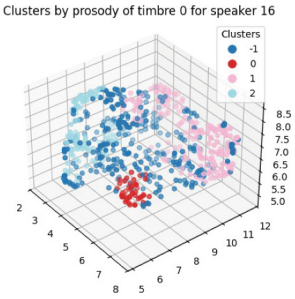
– Style vectors: reduce to 64 dimensions with PCA and then to 3 dimensions with UMAP.
– 'librosa' features: reduce to 9 dimensions with PCA and then to 3 dimensions with UMAP.
– 'emotion2vec' outputs: reduce to 64 dimensions with PCA and then to 3 dimensions with UMAP.

3.3 Feature Combinations

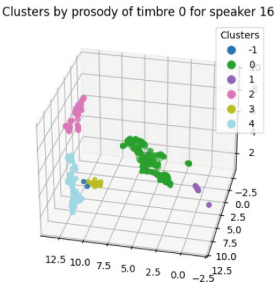
After concatenating all three feature sets the dimensionality is reduced from 9 to 3 (from 9 to 6 with PCA and from 6 to 3 with UMAP). We use clustering of original style vectors as the baseline approach, and clustering of combined features as our final improved approach. As intermediate methods, we cluster the 'librosa' features and the 'emotion2vec' features separately. See Fig. 2.



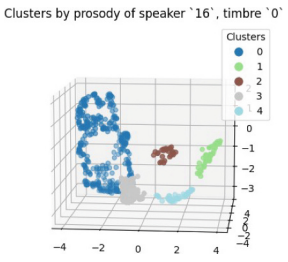
(a) Clustering original style vectors



(b) Clustering 'librosa' features



(c) Clustering features produced by 'emotion2vec'



(d) Clustering combined features

Fig. 2. Clusters formed using different feature combinations generated from speech.

3.4 Feature Emphasis

Four approaches were tested for combining Style TTS 2 style vectors, 'librosa' features, and 'emotion2vec' model outputs, each involving different dimensionality allocation strategies:

- Uniform dimensionality: Equalizing the number of dimensions across style vectors, 'librosa' features, and 'emotion2vec' model outputs to maintain a balanced representation.
- Style vector emphasis: Assigning a higher number of dimensions to style vectors compared to the other features, prioritizing stylistic aspects in the representation.
- 'librosa' feature emphasis: Increasing the dimensionality of 'librosa'-extracted features to enhance the acoustic feature representation.
- 'emotion2vec' model emphasis: Allocating more dimensions to 'emotion2vec' model outputs to strengthen emotional expressiveness in the feature combination.

The uniform dimensionality approach was ultimately selected due to its balanced performance. However, increasing the dimensions allocated to 'librosa' features or 'emotion2vec' model outputs also yielded beneficial results for some speakers. See Fig. 3.

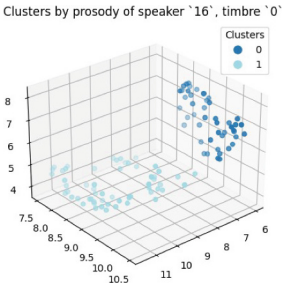
3.5 Evaluation Methods

To evaluate the correctness of the clusters produced by different methods, we employed both qualitative and quantitative approaches. These methods ensured a comprehensive understanding of how well clustering methods captured prosody differences in speech.

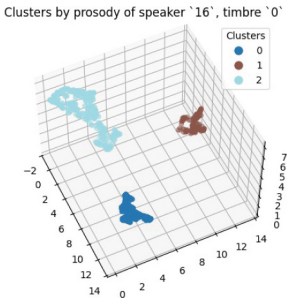
- Visualizations. Clusters were visualized into three-dimensional spaces which allowed us to qualitatively assess:
  - Cluster Separation: Whether distinct clusters are formed for different emotions.
  - Cluster Coherence: Whether data points within a cluster were tightly grouped, indicating that the feature set correctly represents emotions.

By inspecting the visual plots, we could identify patterns and evaluate whether the clustering method effectively differentiated between emotions. Clear separation between clusters representing "amused" and "sleepy" emotions indicated successful clustering, while overlapping clusters suggested potential issues with feature selection or clustering, and dimensionality reduction parameters.

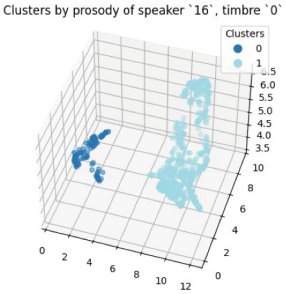
- Emotion Clustering Accuracy with EmoV-DB dataset: The EmoV-DB dataset, which includes emotion labels (e.g., amused, angry, disgusted, neutral, sleepy), was used to quantitatively measure clustering accuracy. The evaluation process involved:



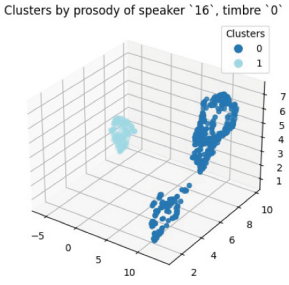
(a) Clustering with style vector emphasis



(b) Clustering with 'librosa' feature emphasis



(c) Clustering with 'emotion2vec' model emphasis



(d) Clustering with uniform dimensionality emphasis

Fig. 3. Clusters formed using features with different source emphasis.

- Cluster Label Distribution Analysis: For each cluster C_k, we calculated the proportion of samples with emotion label e_j as:

$$P(e_j \mid C_k) = \frac{n_{k,j}}{n_k}$$

where $n_{k,j}$ is the number of samples in cluster C_k labeled with emotion e_j, and n_k is the total number of samples in cluster C_k.

- Accuracy Calculation: Clustering accuracy was determined by identifying the dominant emotion label within each cluster and calculating the proportion of samples in that cluster that matched this dominant label. For example, the "amused" emotion achieved 90.32% accuracy within its corresponding cluster. Formally, clustering accuracy was computed as:

$$\text{Accuracy} = \frac{1}{N} \sum_{k=1}^{K} \max_{j}(n_{k,j})$$

where N is the total number of samples, K is the total number of clusters, and $\max_j(n_{k,j})$ gives the count of the dominant emotion in cluster C_k.

This quantitative evaluation provided insights into how well the clustering methods aligned with ground truth labels.

4 Results

After evaluating three different feature sets, we opted to combine them as our final approach, given its superior performance in capturing diverse emotional nuances. This decision was motivated by the results showing that integrating multiple feature types leads to more robust and distinct clustering outcomes.

Our experiments on the EmoV-DB dataset revealed that clustering combined features was the most effective approach, yielding the highest number of well-defined clusters. Emotions such as “angry" (96.43%), “neutral" (99.39%), and “amused" (100.00%) were strongly represented in specific clusters. Additionally, this method improved clustering quality by reducing the number of unclustered data samples by 21.4% and resulted in 25.14% more homogeneous clusters. Despite this, distinguishing between “disgusted" and “sleepiness" remained a challenge, though this issue was confined to only a single cluster similar to the baseline approach. However, since the combined features approach in general generated more clusters, the presence of this issue in just one cluster reflects an improvement in cluster definition compared to the baseline, which produced fewer clusters.

Visualizations from both the EmoV-DB and Hi-Fi TTS datasets supported the effectiveness of combined feature clustering. The resulting clusters were more distinct and cohesive compared to individual feature sets, though some dispersion remained, indicating potential for further refinement. Detailed results are provided in Tables 1, 2, 3, and 4.

Table 1. Clustering results for original style vector features

Style cluster	Amused	Anger	Disgust	Neutral	Sleepiness
0/0	4.09%	-	2.34%	90.06%	3.51%
0/1	98.43%	-	-	-	1.57%
0/2	1.67%	-	45.28%	1.94%	51.11%
1/-1	0.61%	99.39%	-	-	-
2/-1	0.62%	99.38%	-	-	-
3/0	-	-	-	-	100.00%
3/1	-	-	-	-	100.00%
4/-1	100.00%	-	-	-	-
5/0	98.63%	-	-	-	1.37%
5/1	1.30%	-	98.70%	-	-
6/-1	-	-	0.61%	99.39%	-

Table 2. Clustering results for 'librosa' features

Style cluster	Amused	Anger	Disgust	Neutral	Sleepiness
0/0	44.25%	-	38.50%	8.85%	8.41%
0/1	3.38%	-	15.79%	39.85%	40.98%
1/0	-	100.00%	-	-	-
	-	100.00%	-	-	-
	0.62%	99.38%	-	-	-
3/0	-	-	-	-	100.00%
3/1	-	-	-	-	100.00%
4/0	100.00%	-	-	-	-
4/1	100.00%	-	-	-	-
5/0	43.66%	-	56.34%	-	-
5/1	3.57%	-	96.43%	-	-
6/0	-	-	-	100.00%	-
6/1	-	-	-	100.00%	-

Table 3. Clustering results for 'emotion2vec' features

Style cluster	Amused	Anger	Disgust	Neutral	Sleepiness
0/-1	0.62%	99.38%	-	-	-
1/0	-	-	-	-	100.00%
1/1	-	-	-	-	100.00%
2/0	91.57%	-	2.41%	6.02%	-
2/1	2.43%	-	31.98%	21.86%	43.72%
3/-1	-	100.00%	-	-	-
4/-1	-	-	0.61%	99.39%	-
5/0	92.54%	-	7.46%	-	-
5/1	0.90%	-	98.20%	-	0.90%
6/-1	100.00%	-	-	-	-

4.1 EmoV-DB Dataset

For each cluster, we calculated the proportion of samples corresponding to each emotion label.

– *Clustering original style vectors (baseline)*: Some emotions were well-represented within individual clusters, such as "amused" (98.63%), "neutral" (90.06%), and "angry" (99.39%). However, the method struggled to separate "disgusted" and "sleepiness," as their proportions were mixed within clusters. Similar to the combined features approach, this issue appeared in just one cluster, but the baseline method produced fewer clusters overall.

Table 4. Clustering results for combined features

Style cluster	Amused	Anger	Disgust	Neutral	Sleepiness
0/0	26.95%	-	33.54%	1.23%	38.27%
0/1	4.65%	-	2.33%	89.53%	3.49%
1/0	-	100.00%	-	-	-
1/1	-	100.00%	-	-	-
2/0	3.57%	96.43%	-	-	-
2/1	-	100.00%	-	-	-
3/-1	-	-	0.61%	99.39%	-
4/0	-	-	-	-	100.00%
4/1	-	-	-	-	100.00%
4/2	-	-	-	-	100.00%
5/0	100.00%	-	-	-	-
5/1	100.00%	-	-	-	-
5/2	100.00%	-	-	-	-
6/0	-	-	100.00%	-	-
6/1	-	-	100.00%	-	-
7/0	94.44%	-	5.56%	-	-
7/1	100.00%	-	-	-	-

- *Clustering 'librosa' features*: The distribution of emotions varied across clusters, leading to inconsistent results. In one cluster, "amused" and "disgusted" were nearly equally represented (43.66% and 56.34%, respectively), while in another, "disgusted" was dominant (96.43%) with very few "amused" samples (3.57%).
- *Clustering 'emotion2vec' model outputs*: Certain emotions, such as "amused" (92.54%) and "angry" (99.38%), were strongly represented within specific clusters. However, in some cases, emotions like "disgusted" (31.98%), "neutral" (21.86%), and "sleepiness" (43.72%) were more mixed, making them harder to separate.
- *Clustering combined features*: This approach resulted in the highest number of clusters, with many emotions being well-represented in specific ones. For instance, clusters with high proportions of "angry" (96.43%), "neutral" (99.39%), and "amused" (100.00%) were observed. Additionally, this method achieved better clustering quality by reducing the number of unclustered data samples (fewer data assignments to -1 cluster) and producing more cohesive clusters, with some clusters reaching 100% purity for specific emotions. Similar to other methods, this one still struggled to separate "disgusted" and "sleepiness," as their proportions were mixed within one cluster.

4.2 EmoV-DB and Hi-Fi TTS Dataset

- *Clustering original style vectors (see figure 2a)*: The plot shows three clusters. There is a noticeable separation between the brown (0) and light blue (1) clusters. Since only a few points are in the dark blue (-1) cluster, the clustering method clearly assigned most of the data points to a specific prosody group. Overall the clustering method seems fair in distinguishing at least two major groups. However, the clusters seem big and that may indicate that more than one emotion was assigned to one cluster.
- *Clustering 'librosa' features (see figure 2b)*: The red (0), pink (1), and light blue (2) clusters show some separation, suggesting that clustering was able to identify distinct prosodic patterns. However, the dark blue (0) cluster is quite large and dispersed, meaning a significant number of audio samples were not clearly assigned to a specific prosody group. This and the overlapping nature of pink (1) and light blue (2) clusters might suggest that the clustering method struggles to distinguish between certain prosodic variations to effectively capture prosodic patterns.
- *Clustering 'emotion2vec' model outputs (see figure 2c)*: The dark blue (0), brown (1), and light blue (2) clusters exhibit noticeable separation. However, they are not tightly grouped. The clustering algorithm was able to identify distinct emotional patterns from the 'emotion2vec' model outputs.
- *Clustering combined features (see figure 2d)*: The green (1), brown (2), grey (3), and light blue (4) clusters display distinct separations, suggesting that the clustering algorithm effectively identified unique patterns when combining features. The dark blue (0) cluster is large and quite dispersed. However, it also shows its role as a significant and well-identified group with clear borders.

5 Conclusion and Limitations

In this study, we demonstrate that combining StyleTTS 2 style embeddings with additional acoustic-prosodic features and hierarchical HDBSCAN clustering enhances emotional speech synthesis. By leveraging multi-modal feature fusion, we achieve more distinct and expressive emotion clusters compared to previous single-modality approaches. Our results show that the proposed method improves emotion recognition accuracy and provides finer control over synthesized speech prosody and timbre. Notably, the combination of style vectors, 'librosa' -derived acoustic features, and 'emotion2vec' embeddings resulted with superior overall cluster separation and homogenization of said groups compared to baseline. Experiments with the aforementioned method resulted in timbre-prosody clusters such as "angry" (96.43%), "neutral" (99.39%), and "amused" (100.00%), which were strongly represented in specific clusters. Compared to the baseline, our method produced a 200% increase in overall clusters, a 21.4% decrease in unclustered data samples, and a 25.14% increase in homogeneous cluster groups. However, despite these improvements, some emotions such as

"disgusted" and "sleepiness" still exhibited overlap due to inherent similarities in their prosodic characteristics.

Additionally, the performance of our clustering method is influenced by the quality and diversity of the training data. While our approach enhances timbre-prosody disentanglement, generalization across unseen speakers remains a challenge, as the StyleTTS 2 model exhibits speaker bias. Moreover, since our method was primarily evaluated on two datasets, its performance may vary across different speech corpora, languages, or recording conditions.

Future work should focus on developing adaptive feature representations, improving cluster validation methods, exploring cross-lingual generalization to enhance the robustness and scalability of emotion-driven TTS synthesis, and developing a method to automatically label style clusters for efficient emotional TTS.

References

1. Tan, X., et al.: Naturalspeech: end-to-end text-to-speech synthesis with human-level quality. IEEE Trans. Pattern Anal. Mach. Intell. **46**(6), 4234–4245 (2024)
2. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning, pp. 5530–5540. PMLR, (2021)
3. Li, Y.A., Han, C., Raghavan, V., Mischler, G. and Mesgarani, N.: Styletts 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. Adv. Neural Inform. Process. Syst. **36** 19594–19621 (2023)
4. Peng, J., et al.: A survey on speech large language models. arXiv preprint arXiv:2410.18908 (2024)
5. Zeng, Z., Wang, J., Cheng, N., Xiao, J.: Prosody learning mechanism for speech synthesis system without text length limit. arXiv preprint arXiv:2008.05656, 2020
6. Li, T., et al.: DiCLET-TTS: diffusion model based cross-lingual emotion transfer for text-to-speech–a study between English and mandarin. IEEE/ACM Trans. Audio, Speech, Lang. Process. **31**, 3418–3430 (2023)
7. Lameris, H., Mehta, S., Henter, G.E., Gustafson, J., va Székely, É.: Prosody-controllable spontaneous TTS with neural HMMS. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)
8. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779–4783. IEEE, (2018)
9. Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., Lee, S.-W.: Emosphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. arXiv preprint arXiv:2406.07803 (2024)
10. Latif, S., Shahid, A., Qadir, J., et al.: Emotional speech cloning using gans. In: 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 824–828. IEEE (2021)
11. Kaur, N., Singh, P.: Conventional and contemporary approaches used in text to speech synthesis: a review. Artif. Intell. Rev. **56**(7), 5837–5880 (2023)

12. Werchniak, A.: Exploring the application of synthetic audio in training keyword spotters. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7993–7996. IEEE, (2021)
13. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
14. Ma, Z., et al.: emotion2vec: self-supervised pre-training for speech emotion representation. arXiv preprint arXiv:2312.15185 (2023)
15. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
16. McInnes, L., Healy, J., Melville, J.: Umap: uniform manifold approximation and projection for dimension reduction. arxiv. arXiv preprint arXiv:1802.03426, 10, (2018)
17. Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y.: Hi-fi multi-speaker english TTS dataset. arXiv preprint arXiv:2104.01497 (2021)