Emotional Text-To-Speech Based on Mutual-Information-Guided Emotion-Timbre Disentanglement

Jianing Yang*‡, Sheng Li†, Takahiro Shinozaki†, Yuki Saito*, Hiroshi Saruwatari*
*The University of Tokyo, Japan, †Institute of Science Tokyo, Japan.
‡ E-mail: baleyang@g.ecc.u-tokyo.ac.jp

Abstract—**Current emotional Text-To-Speech (TTS) and style transfer methods rely on reference encoders to control global style or emotion vectors, but do not capture nuanced acoustic details of the reference speech. To this end, we propose a novel emotional TTS method that enables fine-grained phoneme-level emotion embedding prediction while disentangling intrinsic attributes of the reference speech. The proposed method employs a style disentanglement method to guide two feature extractors, reducing mutual information between timbre and emotion features, and effectively separating distinct style components from the reference speech. Experimental results demonstrate that our method outperforms baseline TTS systems in generating natural and emotionally rich speech. This work highlights the potential of disentangled and fine-grained representations in advancing the quality and flexibility of emotional TTS systems.**[1]

I. Introduction

Deep learning has significantly advanced Text-To-Speech (TTS) technology, surpassing early statistical models that struggled with naturalness and expressiveness. The introduction of deep neural networks (DNNs) [1] as an acoustic model improved speech fidelity and intelligibility by capturing complex relationships between input text and output speech in TTS. Autoregressive generative models [2]–[4] enabled end-to-end TTS, which represents the mapping from input text to output speech by stacked DNN modules and enhances synthesis quality and robustness toward prosody drift and mis-alignment in long-form synthesis. More recently, non-autoregressive models [5]–[7] have attracted increasing attention. In contrast to autoregressive models, which generate speech frames sequentially, non-autoregressive models predict all output frames in parallel. This parallelization significantly speeds up inference, greatly enhances the efficiency of TTS.

Despite significant advances in naturalness and diversity of synthetic speech by TTS, achieving precise and expressive emotional TTS remains a challenging task. Emotional TTS, particularly in zero-shot settings, aims to generate speech that matches the timbre, emotion, and prosody of a few seconds of reference speech. Traditional methods typically encode the reference speech into a global style vector [8]–[10], which is then fused with the output of a phoneme encoder. Although these approaches effectively capture the overall style, they often struggle to model the phoneme-level variation in emotion and prosody. Moreover, compressing the reference speech into a single global embedding risks losing crucial prosodic details, limiting the expressiveness and control over the synthesized speech.

To address these limitations, we propose a novel emotional TTS method that (i) predicts fine-grained, phoneme-level emotion embeddings, and (ii) disentangles those embeddings from global timbre information through mutual-information minimization. Central to our method is a dedicated Style Encoder, which comprises two parallel extractors: a global Timbre Extractor, and a phoneme-aware Emotion Extractor that aligns reference acoustics with target phonemes to produce an emotion embedding sequence. An unsupervised Mutual Information Neural Estimation (MINE) [11] explicitly pushes the two representations apart, ensuring that the timbre embedding remains speaker-specific information, while the emotion embeddings capture only prosodic nuance, allowing the model to synthesize speech that is simultaneously timbre-consistent and emotionally expressive.

Experiments demonstrate that our method outperforms strong baselines, including Global Style Token(GST) [8], StyleSpeech [9], MIST [12], and DC Comix TTS [13], on both subjective and objective metrics. t-SNE visualizations further reveal well-separated emotion clusters, confirming effective disentanglement. Our results highlight the value of combining phoneme-level emotion modeling with principled feature disentanglement for expressive, high-fidelity emotional TTS.

II. Related Works

TTS has recently progressed from sentence-level style transfer to fine-grained prosody control, yet three technical lines still dominate the literature. Below we summarize each line and highlight the open problem our study tackles.

a) ***Global & hierarchical emotion embeddings****:* [8] first proposed GST—a bank of learnable tokens attended by a reference encoder—to condense an utterance into a single "style vector." Subsequent works refined this idea: StyleSpeech [9] injects the vector into every encoder/decoder block via Style-Adaptive LayerNorm (SALN); Tacotron-GST [10] piles GSTs hierarchically to capture speaking styles spanning

[1] The synthesized audio samples are available at https://baleyang.github.io/emotion-timbre-disentangled-tts/

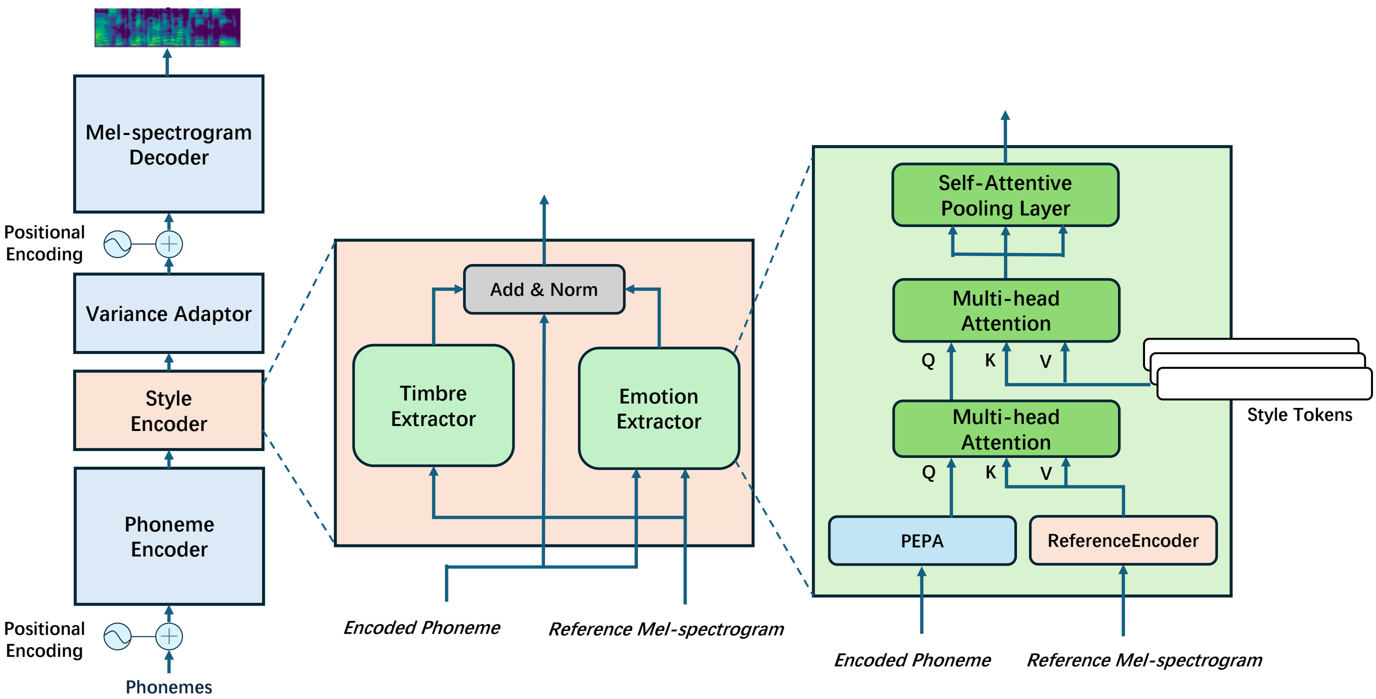

Fig. 1: Overview of the proposed TTS model and its main modules. **Left**: The end-to-end TTS pipeline adopts the FastSpeech 2 backbone; a Style Encoder is inserted after the Phoneme Encoder, followed by the Variance Adaptor and the Mel-spectrogram Decoder. **Center**: Architecture of the Style Encoder, in which separate Timbre and Emotion Extractor are combined through a residual Add & Norm operation. **Right**: Detailed design of the Emotion Extractor, generates phoneme-level emotion embeddings.

words to paragraphs. Although these systems improve expressiveness, their utterance-level embeddings cannot localize phoneme-wise variations and therefore provide only coarse control over emotion and rhythm.

b) ***Prosody modeling via neural codecs:*** Discrete token representations have recently become a focal point in speech modeling research. Neural audio codecs based on vector-quantized variational autoencoders—most notably SoundStream [14] and EnCodec [15]—transform continuous waveforms into sequences of codebook indices, yielding a compact “speech language” that lightweight sequence-to-sequence decoders can handle efficiently. Building on this foundation, several studies have investigated how such tokens can capture prosody for zero-shot TTS. A representative example is DC Comix TTS [13], which tokenizes a reference signal with an EnCodec-style front end and conditions its decoder on a style embedding derived from the resulting code sequence, achieving high-fidelity speech from unseen speakers. However, fixed-rate quantization still blurs micro-prosodic cues—such as subtle emotional nuances and pitch inflections—and the reliance on a global utterance-level embedding limits phoneme-level expressive control.

c) ***Feature disentanglement:*** Most disentanglement studies focus on separating content from a holistic “style” embedding. Typical methods include auxiliary classifiers [16], adversarial objectives [17], and mutual information minimization (MIST; [12]; ProsodySpeech; [18]). Although effective in maintaining intact linguistic information, **they overlook the inter-style entanglement**—timbre, emotion, and local prosody still co-exist in the same vector, making controllable synthesis difficult and hurting zero-shot generalization.

III. Model Architecture

To address the limitations of global style embeddings and leverage MINE for effective speech feature disentanglement, we propose a novel model architecture for emotional TTS that integrates MINE into the style encoding process. Our model introduces two key innovations:

1. Phoneme-level emotion embedding prediction: The model predicts emotional information at the phoneme level while treating timbre as a global feature, enabling TTS that closely matches the style of the reference speech.

2. Effective disentanglement of speech features: By guiding the feature extractor to focus on distinct attributes of the reference mel-spectrogram, we successfully decouple timbre and emotional features, improving the style controllability of synthetic speech.

By combining global timbre prediction with fine-grained emotional modeling and employing MINE for feature disentanglement, our method overcomes the limitations of existing approaches. This improvement significantly enhances the style similarity and expressiveness of the synthetic speech.

A. Model Backbone

The overall architecture of the proposed model is illustrated in Fig. 1. It is based on FastSpeech 2(FS2) [7], consisting of Encoder-Decoder networks with Variance Adaptor, which following the original FS2 method. To enable emotional TTS, we introduce a Style Encoder after the Phoneme Encoder to predict style-specific representations.

Specifically, the Phoneme Encoder consists of four Feed-Forward Transformer (FFT) blocks, while the mel-spectrogram Decoder includes six FFT blocks. The Variance Adaptor comprises a length regulator, pitch predictor, and energy predictor, each implemented using two 1D convolution layers with 256 filters. The structure and design of the Style Encoder are detailed in the next section.

B. Style Encoder

The architecture of our Style Encoder is composed of two parallel modules—a **Timbre Extractor** and an **Emotion Extractor**—that independently leverage the Phoneme Encoder output and a reference mel-spectrogram to derive phoneme-level style information. The central assumption is that timbre remains relatively stable and invariant to specific textual content, whereas emotion and prosody vary significantly with different inputs.

We adopt GST-based method [8] for timbre extraction. Specifically, a Reference Encoder first processes the reference mel-spectrogram to produce an intermediate representation, which is then passed through a style token layer to obtain a global timbre embedding, $\mathbf{F}_{\text{timbre}}$.

For Emotion Extractor, we design a phoneme-aware architecture to capture fine-grained emotional nuances. We begin by encoding the reference mel-spectrogram using the same Reference Encoder employed in the Timbre Extractor.

To bridge the representational gap between the Phoneme Encoder and the Emotion Extractor, we introduce a lightweight Phoneme-Emotion Projection Adapter (PEPA). Implemented as two successive 1-D convolutional layers, PEPA projects the phoneme embeddings generated by the phoneme encoder that is pre-trained exclusively on neutral speech in stage one and kept fixed in stage two (as described in Section IV)—into the prosody-rich acoustic space learned by the Reference Encoder. This projection supplies each phoneme with temporally aligned emotional context, thereby mitigating the mismatch caused by the two-stage training scheme and enabling fine-grained fusion of linguistic and emotional cues.

We then invoke a multi-head cross-attention module that treats the projected phoneme embeddings as queries while using the reference-encoder emotion features as keys and values. For each phoneme position i, the attention matrix yields weights α_{ij} over the emotion sequence e_j; the resulting representation is defined as the weighted sum of α_{ij} and e_j:

$$\tilde{p}_i = \sum_j \alpha_{ij}\, e_j.$$

This operation endows every phoneme with a custom blend of emotional cues proportional to its affinity with each reference frame, producing phoneme-synchronous emotional features that seamlessly fuse linguistic and affective information. Notably, we do *not* include positional encoding for the reference mel-spectrogram features, preventing potential "content leakage" and ensuring robust TTS when the reference mel-spectrogram and target text are mismatched. Following this alignment, we employ another multi-head cross-attention mechanism combined with a style token layer to generate the emotion embedding sequence $\mathbf{F}_{\text{emotion}}$. We then apply a Self-Attentive Pooling Layer [19] to smooth transitions between adjacent phonemes. The final output of these processes is the emotion embedding, $\mathbf{F}_{\text{emotion\_smooth}}$.

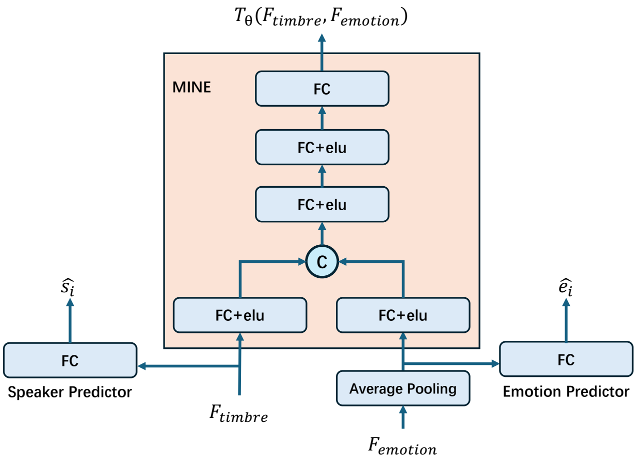

Fig. 2: Proposed architecture for timber-emotion disentanglement based on mutual information minimization

Finally, the Phoneme Encoder output is combined with $\mathbf{F}_{\text{emotion\_smooth}}$ via element-wise addition and then broadcast-summed with $\mathbf{F}_{\text{timbre}}$. After applying layer normalization, the Style Encoder produces its final representation, effectively capturing both global timbre and fine-grained emotional cues.

C. Style Disentanglement

Given the dual-feature extractors (Timbre Extractor and Emotion Extractor) in our model, it is essential to disentangle their respective outputs to ensure that they capture distinct speech attributes. To achieve this, we employ MINE to disentangle the outputs of these two extractors.

Previous studies have attempted to achieve feature disentanglement by minimizing the mutual information between style embeddings. However, these methods face a common issue: the lack of explicit guidance for the disentanglement process. Simply relying on MINE to minimize the mutual information between style embeddings leaves the model with no clear optimization direction, hindering its ability to effectively separate features. To address this, as shown in Fig. 2, we not only minimize the mutual information between the emotion embedding $\mathbf{F}_{\text{emotion}}$ and the timbre embedding $\mathbf{F}_{\text{timbre}}$ using MINE, but also guide the disentanglement by explicitly predicting emotion and speaker labels from $\mathbf{F}_{\text{emotion}}$ and $\mathbf{F}_{\text{timbre}}$, respectively. This approach provides clear optimization objectives, enabling effective emotional speech synthesis.

Specifically, $\mathbf{F}_{\text{emotion}}$ is a sequence of phoneme-level emotion embeddings. We apply average pooling to aggregate this sequence into a global emotion embedding $\mathbf{F}_{\text{emotion\_global}}$. Then, fully connected (FC) layers, named the *Emotion Predictor* and *Epeaker Predictor*, are used to predict the corresponding emotion label and speaker label, providing explicit objectives for optimizing the style extractors.

To estimate and suppress the mutual information between the global-emotion embedding $\mathbf{F}_{\text{emotion\_global}}$ and the timbre embedding $\mathbf{F}_{\text{timbre}}$, we follow the Donsker–Varadhan (DV) variational formulation of the KL divergence [20] that underpins recent works [12], [17], [18], [21]. For any measurable scoring function $T\colon (Y,Z) \mapsto \mathbb{R}$, the DV inequality gives

$$\mathcal{I}(Y,Z) \;\geq\; \hat{\mathcal{I}}_T(Y,Z) \;=\; \mathbb{E}_{P_{Y,Z}}[T] \;-\; \log\bigl(\mathbb{E}_{P_Y \otimes P_Z}[e^T]\bigr).$$

MINE [11] instantiates T as a trainable neural network T_θ and maximizes $\hat{\mathcal{I}}_{T_\theta}$ with respect to the model parameter θ, thereby tightening the lower bound.

In our implementation, T_θ first processes $\mathbf{F}_{\text{emotion\_global}}$ and $\mathbf{F}_{\text{timbre}}$ through two independent fully-connected (FC) layers, each followed by an ELU activation [22]. The resulting vectors are then concatenated and passed to a three-layer FC head whose first two layers again use ELU, while the final layer outputs a scalar score.

During training, the mutual information estimator is optimized by maximizing the negative mutual information $-\hat{\mathcal{I}}_{T_\theta}(Y,Z)$ to capture the dependency between $\mathbf{F}_{\text{emotion\_global}}$ and $\mathbf{F}_{\text{timbre}}$. Simultaneously, the Timbre Extractor and Emotion extractor are updated by minimizing the mutual information $\hat{\mathcal{I}}_{T_\theta}(Y,Z)$, effectively reducing the overlap between the two embeddings. This process ensures that the emotional and timbre features are disentangled, achieving robust style disentanglement.

By combining mutual information minimization with explicit supervision via emotion and speaker labels, our method overcomes the challenges faced by previous methods, enabling the extraction of distinct and independent style attributes for more expressive and controllable emotional TTS.

IV. Implementation

As indicated in [12], [18], [21], obtaining a "clean" encoder whose output phoneme representations do not carry emotion-related information is critical. In the first stage of our method, we therefore train the FastSpeech 2 model *without* the Style Encoder and use only speech samples labeled with the *neutral* emotion category to avoid style-induced variability. The objective in this stage is given by:

$$\mathcal{L}_1 = \mathcal{L}_{\text{recons}} + \lambda_1\,\mathcal{L}_{\text{dur}},$$

where $\mathcal{L}_{\text{recons}}$ denotes the reconstruction loss on the predicted mel-spectrogram, and $\mathcal{L}_{\text{dur}}$ denotes the duration prediction loss. We set $\lambda_1 = 1.0$ throughout our experiments. Once this first-stage training converges, we obtain an encoder that is less sensitive to emotion-related information.

In the second stage, we incorporate the Style Encoder into the model and initialize all encoder parameters using the pre-trained weights from stage one (and freeze these parameters). To ensure the Timbre Extractor and the Emotion Extractor focus on distinct aspects of the reference mel-spectrogram, we adopt a mutual information minimization scheme based on MINE. Specifically, we alternate between updating the TTS model and the MI estimator to encourage the Timbre Extractor and the Emotion Extractor to capture non-overlapping information. We augment our loss with additional pitch and energy terms, as well as classification losses for speaker and emotion. The second-stage objective is defined as:

$$\begin{aligned}\mathcal{L}_2 \;=\; &\mathcal{L}_{\text{recons}} \;+\; \lambda_1\mathcal{L}_{\text{dur}} \;+\; \lambda_2\mathcal{L}_{\text{pitch}}\\ &+\; \lambda_3\mathcal{L}_{\text{energy}} \;+\; \lambda_4\mathcal{L}_{\text{emotion}} \;+\; \lambda_5\mathcal{L}_{\text{speaker}}\\ &+\; \lambda_6\text{ReLU}\bigl(\hat{\mathcal{I}}_{T_\theta}\left(\mathbf{F}_{\text{timbre}}, \mathbf{F}_{\text{emotion\_global}}\right)\bigr),\end{aligned}$$

where $\mathcal{L}_{\text{pitch}}$ and $\mathcal{L}_{\text{energy}}$ measure prediction errors for pitch and energy, respectively, and $\mathcal{L}_{\text{emotion}}$ and $\mathcal{L}_{\text{speaker}}$ are cross-entropy classification losses for emotion and speaker identification, respectively. The term $\hat{\mathcal{I}}_{T_\theta}(\mathbf{F}_{\text{timbre}}, \mathbf{F}_{\text{emotion\_global}})$ is the estimated mutual information, and $-\hat{\mathcal{I}}_{T_\theta}$ is optimized in the MI estimator to capture any correlation between timbre and emotion features. We empirically set $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$, $\lambda_5 = 1.0$, $\lambda_6 = 0.1$, throughout our experiments. By alternating gradient updates between the TTS model and the MI estimator, we encourage the Timbre Extractor and Emotion Extractor to learn disentangled representations.

V. Experiments

A. Dataset

We used emotional speech databases as the main source for our experiments. Specifically, we used the ESD (Emotional Speech Dataset) [23], which contains 350 parallel utterances spoken by 10 native English speakers and 10 native Chinese speakers, covering five distinct emotion categories: neutral, happy, angry, sad, and surprise. In our experiments, we selected the English subset of the dataset. The data was randomly split into 80% for training, 10% for validation, and 10% for testing.

B. Baselines

We conducted a comparison between our model and various baselines.

FS2 + GST. An expressive TTS model that integrates GSTs [8] into the FS2 method to capture diverse speaking styles.

StyleSpeech [9]. An expressive TTS model employing a reference encoder to produce a style embedding, which in turn modulates the output of SALN layers via gain and bias parameters.

FS2 + MIST [12]. A FS2-based model introducing MIST, which reduces the mutual information between the phoneme encoder and the style embedding. This encourages disentanglement of style and content for improved expressive synthesis.

DC Comix TTS [13]. A variant that replaces GST with a reference encoder based on discrete code.

TABLE I: Evaluation results: MOS, MCD, and UAA

Model	MOS(↑)	SMOS(↑)	MCD(↓)	UAA(↑)
FS2 + GST	3.44 ± 0.11	3.12 ± 0.07	8.75 ± 0.22	74.22%
StyleSpeech	2.95 ± 0.11	3.26 ± 0.06	9.09 ± 0.23	82.22%
FS2 + MIST	3.62 ± 0.10	2.89 ± 0.08	8.65 ± 0.22	76.17%
DC Comix TTS	3.59 ± 0.11	2.97 ± 0.08	9.01 ± 0.24	58.79 %
Proposed	$\mathbf{3.63 \pm 0.10}$	$\mathbf{3.41 \pm 0.06}$	$\mathbf{8.23 \pm 0.22}$	**82.42%**

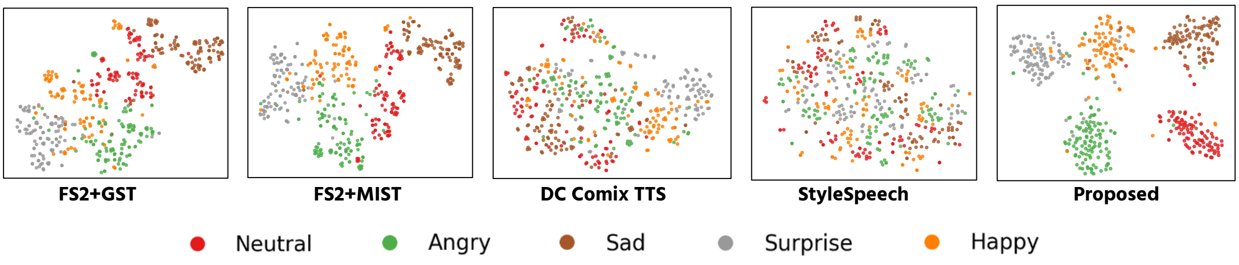

Fig. 3: T-SNE visualization of emotion embeddings

C. Implementation Details

We systematically compared our proposed method with baseline methods, each built on the same FS2 architecture. For a fair comparison, all models followed the same DNN architectures for the encoder, decoder, and variance adaptors. We used the Adam optimizer with β = (0.9, 0.98), $\epsilon = 1 \times 10^{-8}$, and the learning rate ℓ_t follows [1]. The mini-batches contained 64 samples, and all models were trained for 60 k optimization steps. For vocoder, We initially employ the official `UNIVERSAL_V1`[2] version of the pre-trained HiFi-GAN [24], and subsequently perform fine-tuning on the ESD training set to better adapt the vocoder to our data. This adapted vocoder was then used to synthesize the final speech waveforms from the generated mel-spectrograms.

D. Evaluation Metrics

We conducted both objective and subjective evaluations to comprehensively assess our system.

Subjective metrics. Naturalness was evaluated using the classical Mean Opinion Score (MOS), where ten judges rated each utterance on a 1-to-5 scale (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent). To assess emotional or style similarity between reference and synthesized speech, we adopted the Similarity MOS (SMOS), in which the judges scored each utterance pair on a 4-point scale: 1 = Very dissimilar, 2 = Dissimilar, 3 = Similar, 4 = Very similar. All MOS and SMOS results were reported together with their 95% confidence intervals (CI).

Objective metrics. Spectral fidelity was measured by mel-cepstral distortion (MCD), where synthetic and reference mel-spectrograms were first aligned via dynamic time warping (DTW). Expressive adequacy was gauged by an emotion-recognition task: we fine-tuned `openai/whisper-large-v2`[3] on the ESD training split and report the resulting unweighted average accuracy (UAA) on generated speech.

TABLE II: Results of ablation study (emotion/speaker predictors & MINE)

Model	MOS(↑)	SMOS(↑)	MCD(↓)	UAA(↑)
Proposed	$\mathbf{3.62 \pm 0.10}$	$\mathbf{3.54 \pm 0.06}$	$\mathbf{8.23 \pm 0.22}$	**82.42%**
w/o Predictors	3.53 ± 0.10	3.29 ± 0.07	9.71 ± 0.21	56.84%
w/o MINE	3.50 ± 0.09	3.47 ± 0.06	8.59 ± 0.22	76.37%

E. Results

Across all evaluations we prevented “content leakage”—the artificial boost that arose when the reference speech shared lexical content with the synthesis target. Specifically, each reference mel-spectrogram was drawn (with a fixed random seed) from an utterance that matched the speaker and emotion of the target but *differed in text*. Table I *reported* MOS, MCD, and UAA obtained with two pre-trained recognizers. Our model attained naturalness that is statistically on par with the best baseline, while delivering markedly superior style consistency, underscoring its effectiveness at reproducing the intended style without sacrificing perceptual quality.

To visualize the extent of feature disentanglement, Figure 3 *showed* a t-SNE projection of the emotion embeddings. Because our Emotion Extractor *generated* phoneme-level emotion embeddings, we *averaged* them to a single utterance-level embedding before projection. The proposed method *yielded* tight, well-separated clusters for the five emotion categories, whereas embeddings from the strongest baseline *scattered* widely and *overlapped* across classes. This qualitative evidence *corroborated* the quantitative gains and *highlighted* the efficacy of our disentanglement strategy.

F. Ablation Study

To validate the effectiveness of individual components in our proposed model, we conducted ablation experiments, and

[2] https://github.com/jik876/hifi-gan

[3] https://huggingface.co/openai/whisper-large-v2

the results were presented in Table II. In this table, "w/o Predictors" denoted the removal of the Emotion and Speaker Predictors, and their corresponding loss functions were not optimized, whereas "w/o MINE" denoted the exclusion of MINE, meaning the mutual information between the outputs of the timbre and emotion extractors was not minimized. The experimental results demonstrated that incorporating both MINE and the Emotion and Speaker Predictors significantly improved performance. Specifically, the combined use of these components allowed the extractors to better distinguish and capture distinct features from the reference speech. This, in turn, enhanced the overall quality of the synthesized speech.

VI. Conclusions

The study introduced a novel emotional TTS method, enhancing the FS2 architecture with a phoneme-level Emotion Extractor and global Timbre Extractor. To achieve effective disentanglement of style representations, we leveraged a MINE to minimize the mutual information between different feature dimensions. Experimental results demonstrated that our approach consistently outperformed baseline models, highlighting its efficacy.

For future work, we plan to extend our phoneme-level emotion embedding and style disentanglement techniques to multimodal generation and conversational speech dialogue systems, where fine-grained controllability and robust style transfer are equally crucial.

Moreover, we acknowledge that our current backbone is FastSpeech2, which is not state-of-the-art in naturalness and expressivity. As future work, we will port our phoneme-level emotion embedding and disentanglement to diffusion-based and language-model–based TTS backbones.

Acknowledgment: The work was supported by JSPS KAKENHI Grant Numbers 25K21266 and 22K12069, JST Moonshot Grant Number JPMJMS2011 and based on results obtained from a project outsourced by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] A. Vaswani, N. M. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783. DOI: 10.1109/ICASSP.2018.8461368.

[3] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "WaveNet: A Generative Model for Raw Audio," in *Speech Synthesis Workshop*, 2016.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to Human Quality TTS with Transformer," *ArXiv*, vol. abs/1809.08895, 2018.

[5] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *International conference on machine learning*, PMLR, 2020, pp. 7586–7598.

[6] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[7] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," ICLR, 2021.

[8] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *International Conference on Machine Learning*, 2018.

[9] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-StyleSpeech : Multi-Speaker Adaptive Text-to-Speech Generation," *ArXiv*, vol. abs/2106.03153, 2021.

[10] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, *et al.*, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *International Conference on Machine Learning*, 2018.

[11] M. I. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, "Mutual Information Neural Estimation," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 531–540.

[12] T.-y. Hu, A. Shrivastava, O. Tuzel, and C. S. Dhir, "Unsupervised Style and Content Separation by Minimizing Mutual Information for Speech Synthesis," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3267–3271, 2020.

[13] Y. Choi and M.-W. Koo, "DC CoMix TTS: An End-to-End Expressive TTS with Discrete Code Collaborated with Mixer," in *Interspeech*, 2023.

[14] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[15] A. D'efossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *ArXiv*, vol. abs/2210.13438, 2022.

[16] W.-N. Hsu and J. Glass, *Scalable Factorized Hierarchical Variational Autoencoder Training*, 2018. arXiv: 1804.03201.

[17] G. Zhang, S. Qiu, Y. Qin, and T. Lee, "Estimating Mutual Information in Prosody Representation for Emotional Prosody Transfer in Speech Synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5. DOI: 10.1109/ISCSLP49672.2021.9362098.

[18] Y. Yi, L. He, S. Pan, X. Wang, and Y. Xiao, "Prosodyspeech: Towards Advanced Prosody Model for Neural Text-to-Speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7582–7586. DOI: 10.1109/ICASSP43922.2022.9746744.

[19] F. Chen, G. Datta, S. Kundu, and P. A. Beerel, "Self-attentive pooling for efficient deep learning," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3963–3972, 2022.

[20] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time," 1975.

[21] D. Paul, S. Mukherjee, Y. Pantazis, and Y. Stylianou, "A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization," in *Interspeech 2021*, 2021, pp. 3625–3629. DOI: 10.21437/Interspeech.2021-660.

[22] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *arXiv: Learning*, 2015.

[23] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional Voice Conversion: Theory, Databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, 2021.

[24] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *ArXiv*, vol. abs/2010.05646, 2020.