MSE-TTS: Emotion-Controllable Text-to-Speech via Multi-Scale Vision Feature

Xin Dong[1], Yong Yang[1,2], ZhiHao Li[1], Qun Yang[1,*]
[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]Beijing Kedong Electric Power Control System Co., Ltd., Beijing, China
Email: {sx2416152, yangyong2516503, zhihaoli, qun.yang}@nuaa.edu.cn

***Abstract*—Vision-Guided Text-to-Speech aims to generate style-matched speech based on input reference images. Although existing methods can leverage facial images to control speech style, they struggle to fully capture the multi-scale characteristics of facial expressions as well as the fine-grained interactions between emotional features and textual content. To address this challenge, we propose a novel multimodal speech synthesis framework, MSE-TTS. First, we design a Multi-Scale Vision Emotion Encoder to capture richer emotional details. A dual-supervision strategy combining classification and contrastive losses is introduced to enhance feature discriminative power. Second, we introduce an Emotion-Text Fusion Module that employs a bidirectional attention mechanism to achieve deep interaction between emotion vectors and text sequences. Experimental results demonstrate that MSE-TTS outperforms baseline models in both naturalness and emotional expressiveness of synthesized speech. Samples are available at https://mse-tts.github.io/MSE-TTS/.**

***Index Terms*—text to speech, multi modal synthesis, multi-scale feature fusion, emotion control**

I. Introduction

In recent years, speech synthesis technology has made remarkable progress [1]–[6], with research focus gradually shifting toward highly controllable speech generation. Mainstream strategies typically introduce auxiliary modalities to explicitly model stylistic features. Among these, reference-audio-based methods are the most prevalent, such as YourTTS [7] and GenerSpeech [8], which aim to transfer reference audio stylistic features for consistency. Additionally, natural language description-based methods attempt to predict stylistic representations from textual prompts, exemplified by PromptTTS [9] and TSP-TTS [10]. While reference-audio-based methods demand high-quality emotional samples, text-prompt-based approaches are limited by the ambiguity of natural language in expressing fine-grained synthesis intentions. To bridge this gap, vision-guided TTS offers a more intuitive and information-rich interaction paradigm.

Vision-guided speech synthesis controls the style of synthesized speech based on input images, with the workflow shown in Figure 1. In recent years, significant progress has been achieved in this research area. FaceTTS [11] and Face2Speech [12] primarily focus on utilizing images to control speaker identity. MM-TTS [13] and MPE-TTS [14] further attempted to align text, images, and speech into a unified style space for style control. FaceSpeak [15] achieved simultaneous control over identity and emotion by decoupling these features within images. Although extensive research has been devoted to achieving vision control in speech synthesis, two fundamental challenges remain. Firstly, widely used global embeddings capture macro-emotions but lose key micro-expression details critical for emotional expression. Secondly, static vision features cannot finely interact with dynamic text sequences, restricting local emotional fluctuations in synthesized speech.

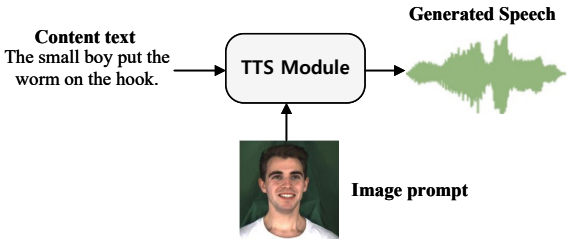


Fig. 1: The pipeline of visual-guided TTS

To tackle these challenges, this paper proposes the MSE-TTS framework. Firstly, to tackle insufficient vision granularity, we design a multi-scale vision emotion encoder with a three-branch architecture to fuse local, intermediate and global features. Dual supervision via classification and InfoNCE contrastive losses is introduced to boost feature discriminability. Secondly, to address fine-grained interactions, we propose an emotion-text fusion module to model deep interactions between emotion and phonemes via bidirectional multi-head attention, enabling dynamic phoneme-level emotion modeling.

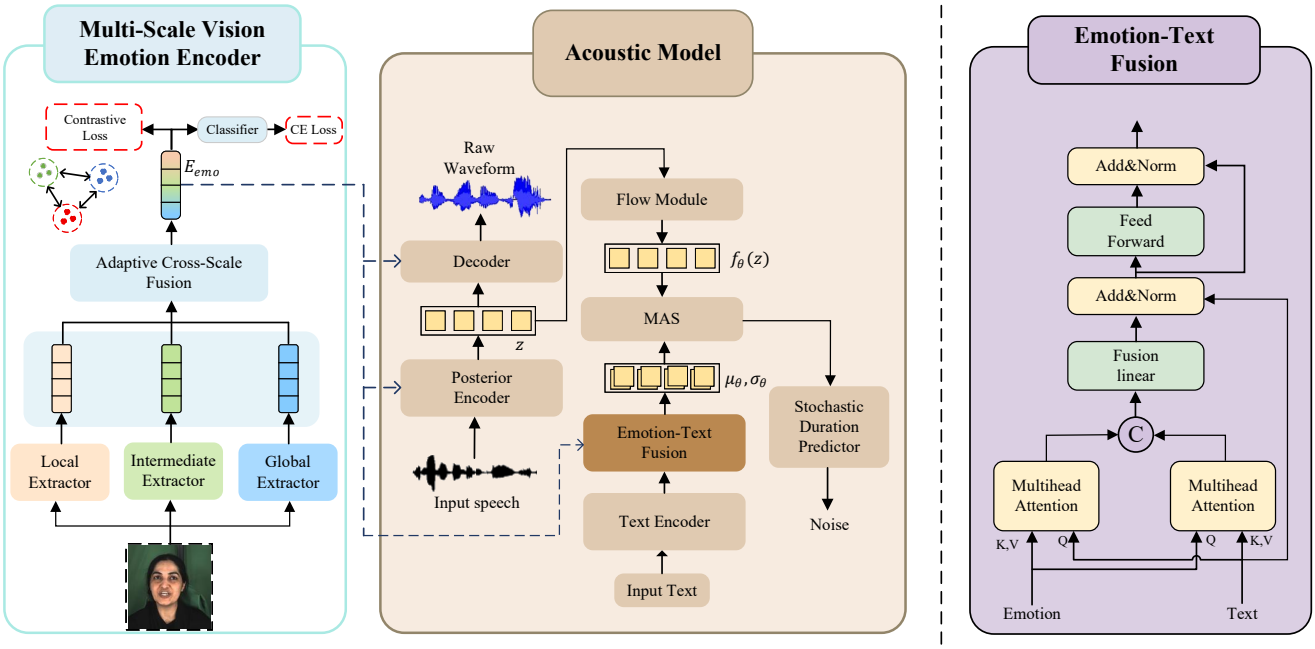
The main contributions of this paper are summarized as follows:

1. We propose MSE-TTS, a vision-driven TTS framework capable of synthesizing high-fidelity, expressive emotional speech from facial images.

2. We design a multi-scale vision emotion encoder that significantly enhances the model's ability to capture subtle facial expressions by integrating multi-scale vision features with dual-supervised loss.

3. We propose a bidirectional attention mechanism for emotion-text fusion that effectively reconciles static vision features with dynamic linguistic content, enabling fine-grained emotion style control in speech synthesis.

*Corresponding Author.



(a) Overall Model Archietecture (b) Emotion-Text Fusion Module

Fig. 2: In (a), the Multi-Scale Visual Emotion Encoder extracts emotional features from the image and feeds them into the Acoustic Model to guide speech synthesis. Within the Acoustic Model, we design an Emotion-Text Fusion module to fuse cross-modal emotional features with textual representations, as illustrated in (b), the © denotes the concatenation operation.

4. Experiments on MEAD-TTS datasets demonstrate that MSE-TTS outperforms baseline models in both objective evaluation metrics and subjective listening tests, generating speech with more nuanced and natural emotional expression.

II. Related Work

A. Expressive Speech Synthesis

In the field of speech synthesis, precise control over style and emotional states is essential for generating high-quality, expressive speech and critical to achieving natural human-computer interaction, thereby motivating extensive research. Mainstream approaches predominantly employ reference audio-driven strategies. GST [16] and VALL-E [17] control style by extracting style embeddings from reference audio. Meta-StyleSpeech [18] and GenerSpeech [8] enhance generalization to unseen styles through meta-learning and multi-scale adapters, respectively. Mega-TTS [19] improves generation robustness by finely decoupling speech attributes. Another class of methods explores the potential of natural language prompts: PromptTTS [9] employs dual encoders to extract stylistic intent from text. PromptTTS 2 [20] further incorporates a variation network to supplement details; while InstructTTS [21] achieves effective decoupling of style and content through mutual information constraints. However, audio-driven approaches remain constrained by the difficulty of acquiring high-quality samples, while text-driven methods struggle to overcome the ambiguity of linguistic descriptions. In contrast, the image modality offers intuitive, easily accessible, and emotionally rich features, providing superior conditions for refined speech synthesis.

B. Vision-Guided Text-to-Speech

Vision-Guided TTS allows a more comprehensive and expressive synthesis process. Prior studies primarily focused on cross-modal biometric consistency. Representative work such as Face2Speech [12] utilize a facial encoder to generate speaker embeddings for controlling the timbre of synthesized speech. FaceTTS [11] further strengthen the identity correspondence between facial images and generated speech through a joint training strategy. Subsequent research gradually expanded into fine-grained audiovisual synchronization and style modeling. Wav2Lip [22] and VisualTTS [23] introduce silent video streams as conditions, focusing on exploring the temporal correlation between visual lip movements and speech content. Regarding style control, FaceSpeak [15] attempts to decouple identity and emotional features through decoupled representation learning to manipulate speech emotion. MM-TTS [13] and MPE-TTS [14] constructs a multimodal prompt encoder aiming to map images, text, and audio into a unified style space, enabling style-controllable TTS based on multimodal information. Although these approaches have expanded the boundaries of vision-driven TTS, significant technical bottlenecks persist. Firstly, static global embeddings lose critical micro-expression details. Secondly, static vision features fail to interact with dynamic text sequences in a

fine-grained manner, resulting in emotionally flat synthesized speech.

To mitigate the aforementioned limitations, we introduce MSE-TTS, a framework that couples a multi-scale visual encoder with an emotion-text fusion mechanism, enabling hierarchical micro-expression modeling and fine-grained cross-modal alignment for emotionally expressive speech synthesis.

III. Method

A. Overall Architecture

The overall architecture of the proposed model is illustrated in Figure 2. MSE-TTS comprises two primary sub-modules: (1) a Multi-Scale Vision Emotion Encoder that extracts hierarchical vision features; and (2) a acoustic module built upon VITS, incorporating our proposed Emotion-Text Fusion Module to synthesize the speech waveform.

B. Multi-Scale Vision Emotion Encoder

Existing methods typically encode images into a single global vector, leading to the loss of fine-grained visual information. To overcome this limitation, we propose a multi-scale vision feature extraction module based on a three-stream parallel architecture. By employing differentiated receptive field constraints, our approach achieves comprehensive parsing of input images across three distinct levels: texture, structure, and semantics.

Multi-scale Feature Extraction. The multi-scale feature extraction module comprises three parallel sub-networks operating at local, intermediate and global scales to extract vision features. Specifically, we design a Local Extractor employing a lightweight small-kernel architecture to capture high-frequency texture details and micro-expression information, yielding F_{local}. We construct a Intermediate Extractor that utilizes medium receptive fields to model the geometric spatial relationships among facial components, thereby bridging the gap between low-level textures and high-level semantics to generate $F_{intermediate}$. Furthermore, we devise a Global Extractor based on deep networks with global receptive fields to extract facial contours and macro-expression semantics, establishing the holistic emotional tone to obtain F_{global}. The detailed structures of each module are provided in the Experiments section.

Adaptive Cross-Scale Fusion. To tackle the distribution heterogeneity of emotional features across samples, we design an adaptive cross-scale fusion module for instance-level dynamic feature aggregation. Specifically, an MLP-based saliency predictor projects concatenated multi-scale features into the weight space to adaptively learning the importance of each scale:

$$W = Softmax(MLP(F_{local} \oplus F_{intermediate} \oplus F_{global})) \quad (1)$$

where $\oplus$ denotes feature concatenation, and $W = [w_1, w_2, w_3]$ represents the normalized adaptive weight coefficients. Subsequently, the final sentiment representation is generated through weighted aggregation:

$$E_{emo} = w_1 F_{local} + w_2 F_{intermediate} + w_3 F_{global} \quad (2)$$

The resulting E_{emo} adaptively amplifies the most discriminative local details pertinent to the current emotional expression while maintaining global semantic consistency, thereby achieving complementary enhancement across scales.

C. Emotion-Text Fusion Module

Fine-grained emotional speech synthesis necessitates deep bidirectional interaction between static emotion cues and dynamic linguistic content. Nevertheless, conventional approaches typically treat emotional features as global static conditions directly injected into the synthesis network, resulting in limited fine-grained controllability. To this end, we design an emotion-text fusion module, as illustrated in Figure 2(b), which employs a dual-path cross-attention architecture to model cross-modal dependencies. This architecture simultaneously enables top-down emotional modulation of phoneme distributions and bottom-up semantic constraints on emotional expression, thereby achieving deep fusion of cross-modal emotional features and linguistic representations.

Specifically, let E_{emo} denote the emotion embedding extracted by the multi-scale visual encoder, and $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_L\}$ represent the hidden state sequence output by the text encoder, where L indicates the phoneme sequence length. Prior to entering the fusion module, we expand the emotional representation E_{emo} along the temporal dimension and inject positional encoding to obtain $\mathbf{E}$. We achieve deep fusion of emotional features and textual features through the Emotion-Text Fusion Module, as detailed below.

Emotion-guided Text Attention. We treat emotional features as queries and textual features as keys and values. The attention output is computed as follows:

$$H_{E \to T} = Attention(E, T, T) \quad (3)$$

$\mathbf{H}_{E \to T}$ denotes the textual features queried by the emotional features. This constructs a text saliency map from the emotion perspective, enabling the model to automatically identify phoneme units most significantly influenced by emotional intensity, thereby addressing the question of "where emotions should be emphasized."

Text-guided Emotion Attention. To achieve phoneme-level emotional alignment, we treat text features as queries and emotional features as keys and values. The computation is formulated as follows:

$$H_{T \to E} = Attention(T, E, E) \quad (4)$$

$\mathbf{H}_{T \to E}$ denotes the emotion attended by textual features. This enables phonemes to dynamically retrieve the most compatible emotional cues from emotional features based on their semantic context, thereby addressing the question of "what emotion this phoneme should express."

Fusion and Refinement. The outputs from both attention mechanisms are subsequently processed to yield the final

output H_{fused}. Formally, the overall computational procedure is defined as follows:

$$H_1 = LayerNorm\left([H_{E\to T} \oplus H_{T\to E}]W^O + h_{res}\right) \quad (5)$$

where $\mathbf{h}_{\text{res}}$ denotes the residual connection from the input text features, W^O denotes the output projection matrix, $\oplus$ indicates feature concatenation, and $LayerNorm$ denotes layer normalization. This yields a preliminary fused state with bidirectional alignment H_1.

Subsequently, we obtained refined representations of cross-modal fused features H_{ffn} as follows:

$$H_{ffn} = FFN\left(H_1\right) \quad (6)$$

where FFN represents the feed-forward network.

Finally, H_{fused} is generated via the following equation, which enables thorough fusion of emotional representations and textual.

$$H_{fused} = LayerNorm\left(H_{ffn} + H_1\right) \quad (7)$$

D. Acoustic Model

We utilize VITS as the foundational acoustic model. VITS is fundamentally a conditional variational autoencoder that integrates normalizing flows [24] with adversarial training. Its core architecture comprises a posterior encoder, a prior encoder (consisting of a text encoder and flow module), and a HiFi-GAN-based decoder. Practically, we add an emotion-text fusion module after the text encoder and inject emotion information extracted from images into the prior encoder, decoder, and emotion-text fusion module respectively.

E. Loss function

Dual Supervision Loss. To ensure that the extracted emotional embeddings E_{emo} exhibit both high discriminability and semantic consistency in the latent space, we propose a dual-supervision training strategy combining cross-entropy loss with InfoNCE contrastive loss.

We first introduce the cross-entropy loss L_{cls} to provide explicit supervision through ground-truth labels, driving the encoder to learn class-discriminative high-level semantic features.

However, classification supervision alone is insufficient to ensure compact intra-class distributions. So, we further incorporate an InfoNCE-based contrastive loss (L_{cont}) to enhance the quality of emotional representations by optimizing the geometric structure of the feature space. It maximizes the similarity between intra-class samples while facilitating the separation of inter-class features. Thus, it can drive the encoder to strip away redundant noise unrelated to emotion and distill more robust essential emotion features. Specifically, given a batch of embedding vectors $\{z_i\}_{i=1}^N$ and their corresponding labels $\{y_i\}_{i=1}^N$, the contrastive loss is defined as:

$$L_{cont} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp(sim(z_i,z_j)/\tau)}{\sum_{k=1}^{N}\exp(sim(z_i,z_k)/\tau)} \quad (8)$$

where N denotes the batch size, τ represents the temperature parameter, $P(i)$ indicates the set of positive samples for sample i (those sharing the same label but excluding the sample itself), $|P(i)|$ denotes the number of positive samples.

Therefore, the dual supervision loss is defined as follows:

$$L_{emo} = \alpha L_{cls} + \beta L_{cont} \quad (9)$$

where α and β represent balance coefficients. In the experiments, we set α = 1 and β = 0.5.

Final Optimization Objective. The total loss function of MSE-TTS is defined as the combination of the dual-supervised emotional loss and the base VITS loss:

$$L = L_{vits} + L_{emo} \quad (10)$$

where L_{vits} denotes the loss term from the original VITS.

IV. EXPERIMENTS

A. Dataset

Following MM-TTS [13], we conduct experiments on the MEAD-TTS [13] dataset. To the best of our knowledge, it is currently the only publicly available multimodal emotional speech dataset compatible with our task. The dataset comprises 31,055 aligned tri-modal (Audio-Vision-Text) samples. It comprises approximately 36 hours of audio sampled at 16 kHz from 47 speakers (27 male/20 female). The dataset covers 8 typical emotion categories, and its rich multimodal annotations enable models to effectively learn complex emotion and style transfer capabilities.

B. Implementation Details

Model Structure. The multi-scale visual encoder extracts features and projects them into a unified 512-dimensional space: the local extractor employs two stacked 3×3 convolutional layers; the intermediate extractor adopts a cascade of two 5×5 and one 3×3 convolutions, supplemented by 2×2 pooling; the global extractor uses large-kernel convolutions like 7×7, 5×5 and 4×4 pooling to capture context. The emotion-text fusion module is constructed upon the multi-head attention mechanism ($N_{head} = 2$).

Training. The model was trained on two NVIDIA GeForce RTX 4090 GPUs with a batch size of 32. Training lasted for 300 epochs. We employ the AdamW optimizer with hyperparameters β_1 = 0.8 and β_2 = 0.99. The initial learning rate was set to 1×10^{-4}, and an Exponential Learning Rate Scheduler was implemented to dynamically adjust the learning rate during training.

C. Baselines and Evalution Metrics

To validate the effectiveness of the proposed method, we conducted comparative experiments between our model and the following representative systems:

• GT (Ground Truth): Original authentic speech recordings from the dataset.

• VITS: An advanced end-to-end TTS model employing conditional variational autoencoders with adversarial training.

• FaceTTS: This model jointly trains a biometric system with a TTS module, utilizing facial images to guide synthesized speech.

• FaceSpeak: This approach separates identity and emotion representations from facial images, thereby controlling the style of synthesized speech.

All models are trained on the MEAD-TTS dataset to ensure a fair comparison. To comprehensively evaluate the model, we employ a combined subjective-objective assessment strategy.

Subjective evaluation. We employ Mean Opinion Score (MOS) and Emotion Similarity Mean Opinion Score (ESMOS), with all scores reported as 95% confidence intervals, supplemented by UTMOS for automatic quality prediction.

Objective evaluation. We employ MCD [25], FFE [26], GPE, VDE [27], WER and Emo-Acc (emotion accuracy) for evaluation. Specifically, WER is calculated using Whisper-Large [28], while Emo-Acc is computed via a fine-tuned Wav2Vec 2.0 [29] classifier to quantify the emotion category accuracy of the synthesized speech.

D. Experiment Result

Subjective Evaluation. As shown in Table I, the subjective evaluation results validate the effectiveness of MSE-TTS in speech synthesis tasks. Specifically, the ESMOS metric indicates that our approach exhibits significant superiority in emotional speech synthesis. The MOS and UTMOS scores further confirm that the synthesized speech achieves high levels of naturalness and fluency.

TABLE I: MOS results with 95% confidence interval

Model	*UTMOS*↑	*MOS*↑	*ESMOS*↑
GT	2.49	4.537 ± 0.247	-
VITS	2.32	3.243 ± 0.362	2.712 ± 0.644
Facespeak	2.13	3.222 ± 0.388	3.236 ± 0.446
Facetts	1.87	3.061 ± 0.458	2.970 ± 0.429
ours	**2.32**	**3.938 ± 0.341**	**4.023 ± 0.329**

Objective Evaluation. As shown in Table II, MSE-TTS comprehensively outperforms baseline models in objective evaluation. Emo-Acc, WER, and MCD all achieve optimal performance. Meanwhile, the significant reduction in FFE and GPE indicates high fundamental frequency prediction accuracy. Although VDE exhibits slight fluctuations, the overall prosodic metrics still validate the effectiveness of the model.

TABLE II: Objective Evaluation results

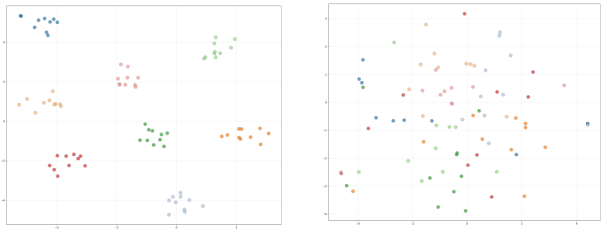
Model	*WER*↓	*EMO-ACC*↑	*MCD*↓	*FFE*↓	*GPE*↓	*VDE*↓
GT	9.99	0.84	-	-	-	-
VITS	21.32	0.335	5.42	0.1374	0.2323	0.0093
Facespeak	24.33	0.49	5.10	0.1047	0.1804	**0.0089**
Facetts	17.40	0.485	4.74	0.1012	0.1661	0.0124
ours	**15.34**	**0.63**	**4.41**	**0.0795**	**0.1336**	0.0091

E. Ablation Studies

As shown in Table III, the ablation studies validate the efficacy of critical modules. Removal of either the emotion-text fusion or contrastive learning degrades performance across all metrics. The T-SNE visualization results presented in Figure 3 further validate the effectiveness of contrastive loss. The model incorporating contrastive loss yields discriminative compact clusters (Figure 3(a)) versus scattered distributions upon ablation (Figure 3(b)), achieving optimal performance.

TABLE III: Ablation Study results

Model	*WER*↓	*EMO*↑	*MCD*↓	*FFE*↓	*GPE*↓	*VDE*↓
ours	**15.34**	**0.63**	**4.41**	**0.0795**	**0.1336**	**0.0091**
w/o contrastive loss	25.10	0.415	4.93	0.0959	0.1710	0.0093
w/o Fusion Module	22.85	0.45	4.80	0.0925	0.1620	0.0096



(a) With Contrastive Loss (b) Without Contrastive loss

Fig. 3: T-SNE visualization of emotion embeddings. Each distinct color denotes a specific emotional label.

F. Scale Weight Distribution Analysis

Figure 4 illustrates the weight distributions of multi-scale features across different emotion categories. The Global scale predominates consistently (weights>0.44), indicating macroscopic facial structures govern emotional representation. Local weights exhibit adaptive variations across categories, validating their capacity for fine-grained expression discrimination. The Intermediate scale maintains stable low weights ($0.10 \sim 0.21$), acting primarily as semantic bridges. Notably, all scales sustain baseline contributions above 0.1, providing distinct yet complementary representational information and thereby validating the efficacy of the multi-scale architecture.

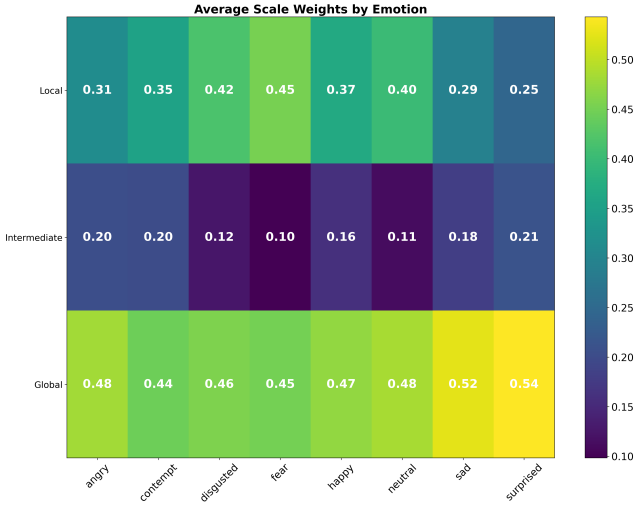


Fig. 4: Scale Weight Distribution.

V. CONCLUSION

In this paper, we present MSE-TTS, a novel multimodal TTS framework. Firstly, to address the coarse-grained visual representations in existing methods, we propose a multi-scale visual emotion encoder that captures features ranging from subtle micro-expressions to holistic facial semantics. Secondly, we introduce an emotion-text fusion module incorporating bidirectional attention mechanisms to effectively model fine-grained interactions between static visual cues and dynamic linguistic content. Comprehensive experiments on the MEAD-TTS dataset demonstrate the superiority of our approach.

REFERENCES

[1] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proceedings of the Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3165–3174.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 5530–5540.

[3] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 4700–4709.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[6] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 8599–8608.

[7] E. Casanova, J. Weber, C. D. Shulby, A. C. Júnior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 2709–2720.

[8] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *Proceedings of the Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[9] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] D. Seong, H. Lee, and J. Chang, "TSP-TTS: text-based style predictor with residual vector quantization for expressive text-to-speech," in *Proceedings of the 25th Annual Conference of the International Speech Communication Association (Interspeech)*, 2024.

[11] J. Lee, J. S. Chung, and S. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 1321–1325.

[13] W. Guan, Y. Li, T. Li, H. Huang, F. Wang, J. Lin, L. Huang, L. Li, and Q. Hong, "MM-TTS: multi-modal prompt based style transfer for expressive text-to-speech synthesis," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2024, pp. 18 117–18 125.

[14] Z. Wu, Y. Kang, S. Cao, L. Ma, Q. Li, and Q. Yang, "MPE-TTS: customized emotion zero-shot text-to-speech using multi-modal prompt," in *Proceedings of the 26th Annual Conference of the International Speech Communication Association (Interspeech)*, 2025.

[15] T. Zhang, J. Zhang, J. Wang, X. Qian, and X. Yin, "Facespeak: Expressive and high-quality speech synthesis from human portraits of different styles," in *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2025, pp. 25 922–25 930.

[16] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 5167–5176.

[17] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.

[18] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 7748–7759.

[19] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, Z. Ma, and Z. Zhao, "Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias," *arXiv preprint arXiv:2306.03509*, 2023.

[20] Y. Leng, Z. Guo, K. Shen, Z. Ju, X. Tan, E. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, L. He, X. Li, S. Zhao, T. Qin, and J. Bian, "Prompttts 2: Describing and generating voices with text prompt," in *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

[21] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "Instructtts: Modelling expressive TTS in discrete latent space with natural language style prompt," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 2913–2925, 2024.

[22] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, 2020, pp. 484–492.

[23] J. Lu, B. Sisman, R. Liu, M. Zhang, and H. Li, "Visualtts: TTS with accurate lip-speech synchronization for automatic voice over," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8032–8036.

[24] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 37, 2015, pp. 1530–1538.

[25] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of the IEEE pacific rim conference on Communications, Computers and Signal Processing (PACRIM)*, 1993, pp. 125–128.

[26] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3969–3972.

[27] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Communication*, vol. 50, no. 3, pp. 203–214, 2008.

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 28 492–28 518.

[29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 12 449–12 460.