Prompt-Unseen-Emotion: Mixed Emotional Speech Synthesis with Prompt-LLM Contextual Knowledge

Xiaoxue Gao, *Senior Member, IEEE*, Huayun Zhang and Nancy F. Chen, *Senior Member, IEEE*

***Abstract*—Existing expressive text-to-speech (TTS) systems primarily model a limited set of categorical emotions, whereas human conversations extend far beyond these predefined emotions, making it essential to explore more diverse emotional speech generation for more natural interactions. To bridge this gap, this paper proposes a novel prompt-unseen-emotion (*PUE*) approach to generate unseen emotional speech via emotion-guided prompt learning. *PUE* is trained utilizing an LLM-TTS architecture to ensure emotional consistency between categorical emotion-relevant prompts and emotional speech, allowing the model to quantitatively capture different emotion weightings per utterance. During inference, mixed emotional speech can be generated by flexibly adjusting emotion proportions and leveraging LLM contextual knowledge, enabling the model to quantify different emotional styles. Our proposed *PUE* successfully facilitates expressive speech synthesis of unseen emotions in a zero-shot setting.**

***Index Terms*—LLM, speech synthesis, emotion, zero-shot TTS.**

I. Introduction

HUMANS naturally produce speech with a wide range of emotional variations [1]–[6]. Expressive speech synthesis seeks to capture this diversity by converting text into speech that not only sounds human but also conveys a genuine emotional tone [7]–[12]. To generate genuinely realistic expressive speech, expressive text-to-speech (TTS) systems must account for a range of factors that extend beyond the textual content alone [13]–[17].

These include subtle ways in which multiple emotions are expressed through the intricate interplay of human emotional characteristics and the robustness of the model [18]–[20], especially given that humans are capable of experiencing approximately 34,000 distinct emotions and even multiple emotional states simultaneously [21]–[23]. For instance, humans naturally feel mixed emotions in their daily life, such as attending a child's wedding can evoke both happiness seeing their child embark on a new journey and a poignant sense of loss [24]. This underscores the importance of emotional TTS models that can accurately simulate the co-occurrence of diverse emotional states to reflect human experiences [24].

Xiaoxue Gao, Huayun Zhang and Nancy F. Chen are with the Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mails: Gao_Xiaoxue@a-star.edu.sg; zhang_huayun@a-star.edu.sg and nancy_chen@a-star.edu.sg). This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004), by A*STAR under its Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) and A*STAR 2024 (R24I6IR136), and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme (DesCartes). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

Expressive speech synthesis has been extensively studied relying on the data-driven training approaches, including Tacotron [7], [10], [17], [25], long-short-term memory network [26], [27], FastSpeech [8]–[10], [12], [16], VITS architecture [28], diffusion models [11], flow-matching models [15] and large language model (LLM)-based systems [29], [30], [30], [31]. These data-driven approaches rely on learning the temporal emotional structure of a fixed set of speech emotions, typically limited to at most eight categories in existing emotional TTS databases. As a result, their expressiveness is constrained, as the synthesized speech is restricted to reproducing only the emotion types present in the training.

Although these methods excel in generating emotional speech for specific categories such as anger, happiness, sadness, and surprise, the generation of unseen emotional expressions, such as disappointment, remains an open challenge [15], [26]–[28]. Existing methods are still struggling to capture complex unseen emotional styles needed for real-world applications [7], [10], [17], [25]. One particularly challenging aspect is generating mixed-emotion speech, which requires modeling intricate combinations of emotional expressions.

In line with this, the emotion wheel theory proposes that all emotions stem from eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and happiness [32]. The remaining 34,000 emotional states are viewed as mixed or derivative forms of these foundational categories [32]. For instance, disappointment can be conceptualized as a blend of surprise and sadness [24], [33]. The only prior work for mixed emotional TTS employs a VITS-based framework with a relative scheme [24], which, although constrained in speech quality, provides valuable insights for this study. In contrast, other mixed-emotion generation methods focus on emotional voice conversion [34], which takes speech as input rather than text, making it fundamentally different from emotional TTS while still providing relevant insights for this work.

Motivated by the success of in-context learning abilities in LLMs [35]–[38] and the need to model more diverse emotional styles that go beyond predefined emotional categories [24], we propose a prompt-unseen-emotion (*PUE*) approach with emotion-guided prompt learning technique to capture the nuanced characteristics of mixed emotional expressions. Emotion-guided prompt learning is designed to capture and quantify different emotional styles by modeling the emotional proportions embedded within the prompts. *PUE* is trained using an LLM-based TTS architecture, which conditions the emotional context on speech generation to maintain emotional consistency between the contextual prompts and the corresponding speech data. By incorporating various emotion

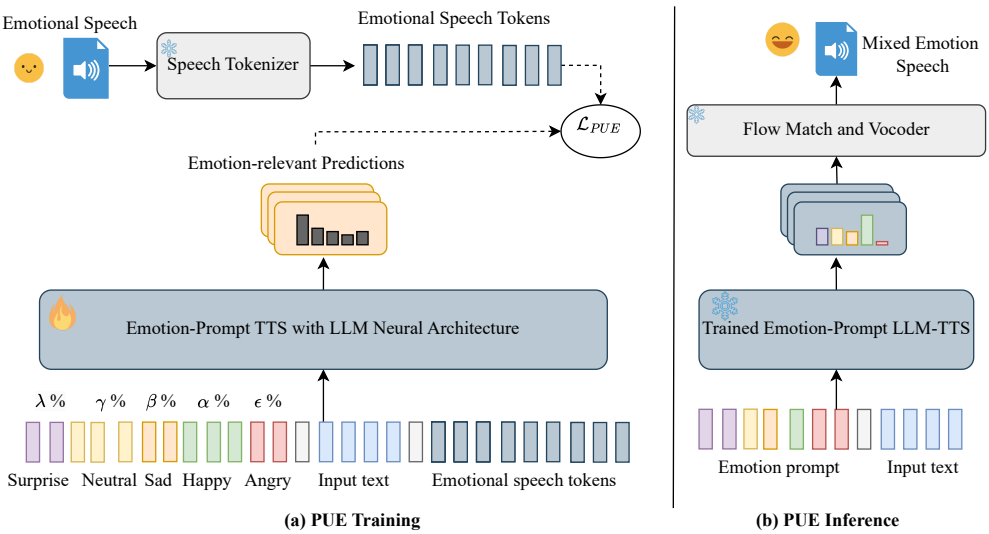

Fig. 1. An overview of the proposed *PUE* approach: (a) the *PUE* training process and (b) the *PUE* inference process.

style descriptions into the emotion-relevant contexts, the *PUE* leverages the LLM's in-context learning capabilities to learn the distribution of different emotional proportions for each utterance. This enables the generation of unseen emotions by adjusting the emotion proportions in the prompts during inference, offering flexibility in emotional expression.

The main contributions of this paper are as follows: (a) the proposed *PUE* method is the first to leverage the in-context learning capabilities of LLMs through prompting for the purpose of expressive speech generation of unseen emotion types that are challenging to obtain in real-world datasets; (b) *PUE* facilitates zero-shot expressive speech synthesis by simply modifying emotion-relevant prompts during inference, enabling mixed-emotion speech synthesis without the need for additional training on new emotion categories; and (c) Extensive objective and subjective evaluations demonstrate that *PUE* outperforms baseline methods, achieving superior mixed-emotion speech generation.

II. Methodology-PUE

We propose *PUE*, an expressive speech synthesis method that incorporates emotion-guided prompt learning within an LLM-based TTS architecture (Fig. 1).

A. Overview

PUE generates emotional speech conditioned on both textual content and specified emotional inputs, with a particular focus on synthesizing mixed emotions that are not encountered during training. This method combines emotion-guided prompts with an LLM-based TTS network, optimized to maximize the likelihood of producing a speech token sequence that accurately reflects the emotional cues provided by the emotion-guided instructional data, as illustrated in Fig. 1.

B. Emotion-guided Prompt Learning

We propose leveraging emotion-guided prompt learning with an LLM-TTS neural architecture θ to take advantage of the instruction-following abilities and in-context learning capabilities from LLMs using paired emotional textual and speech data. Each training instance is structured as:

$$d_i \in D_{\text{e}} = EP\texttt{<EOP>}t_i\texttt{<T>}s_i\texttt{<E>} \quad (1)$$

where EP, t_i, s_i, <EOP>, and <T> represent the emotion-guided prompts, the textual token sequence, the emotional speech token sequence corresponding to the current emotion, the-end-of-prompt token marking the end of emotion-guided knowledge trigger, the turning token from text to speech, and the-end-of-sequence token, respectively. To enable mixed-emotion contextual prompt learning, we propose conditioning the LLM-TTS model with diverse emotion styles by formulating the emotion-guided prompt EP as "A man/woman speaks an utterance with α percent happy emotion, β percent sad emotion, γ percent neutral emotion, ϵ percent angry emotion, and λ percent surprise emotion." The LLM-TTS architecture contains a speech tokenizer and an emotion-prompted LLM-TTS model with a text encoder and an auto-regressive LLM decoder. The tokenizer encodes speech into discrete emotional token sequences, while the auto-regressive LLM decoder predicts the emotion-relevant probability distribution of emotional speech tokens across varying emotion styles.

Emotion-guided prompts are automatically constructed by assigning different scalars to the emotion-relevance parameters based on the emotional category in the dataset. This facilitates the implicit modeling of mixed-emotion generation through the instruction-following capabilities inherent to large language models. Through this mechanism, the proposed *PUE* is contextually guided with the weightings of various emotion proportions per utterance in a quantitative manner, allowing finer control over the emotional composition of each utterance. Given the high cost of collecting and annotating mixed-emotion speech data, our *PUE* offers substantial flexibility for zero-shot expressive speech generation without requiring such data during training.

C. PUE Training Objective

Motivated by the success of label smoothing Kullback-Leibler loss in speech generation [39], we employ an emotion-guided Kullback-Leibler loss to minimize the divergence between the emotional probability distribution predicted by θ, denoted as P_θ, and the target emotional distribution P:

$$\mathcal{L}_{PUE} = KL(P_\theta||P),$$
$$KL(P_\theta||P) = E_{d_i \sim D_{\text{e}}}\left[p(s_i|EP,t_i)\log\frac{p(s_i|EP,t_i)}{p_\theta(s_i|EP,t_i)}\right] \quad (2)$$

Here, the emotional speech tokens s_i are first extracted from a pretrained speech tokenizer, and θ learns to predict the

TABLE I
OBJECTIVE EVALUATION RESULTS COMPARISON OF THE PROPOSED *PUE* WITH BASELINES ON SPEECH INTELLIGIBILITY IN TERMS OF SPEECH INTELLIGENCE THROUGH WER.

TTS Models	**mix** [24]	**cosy** [39]	***PUE***
original test	32.59	51.85	**26.67**
surprise + 30% others	32.59	47.41	**25.19**
surprise + 60% others	31.11	42.22	**25.93**
surprise + 90% others	27.41	39.26	**25.93**

probability distribution of these tokens conditioned on the emotion-guided prompts EP and text t_i. The target distribution P corresponds to the emotion proportions specified in the prompts. Although the training data only contain single-emotion utterances (e.g., 100% angry, 0% happy, 0% sad), the model is explicitly informed of both present and absent proportions through the prompts. This allows θ to associate numerical proportions with emotional styles and, during inference, generalize to mixed-emotion rendering by adjusting the prompt weights. In this manner, θ learns to produce the mixed-emotion speech token sequences that are consistent with the emotion-relevant contexts provided in the input instruction, ensuring that the generated expressive speech captures the weighting of emotion mixture contexts as indicated by EP. During inference, emotion-guided prompts can be adjusted by assigning different values to α, β, γ, ϵ and λ, which define the proportions of relative emotion styles, thus facilitating the generation of mixed-emotion speech.

III. EXPERIMENTS

We employ Cosyvoice-300M-Instruct model (cosy) [39] and the Mixed Emotion model (mix) [24] as strong baselines where no speech prompts are employed. For both cosy and our proposed *PUE* model, a pretrained flow-matching model and a pretrained HiFi-GAN vocoder [39] are used during inference. Man and woman in Equation (2) are decided by the gender of the desired speech output. *PUE* is trained for one epoch using dynamic batching on four GPUs, whose LLM-based TTS model, the speech tokenizer, and the text encoder are initialized from CosyVoice, maintaining the same architectural configurations. Following [24] during inference, we generate zero-shot mixed-emotional speech by combining a primary emotion (surprise) with three other emotions (happy, angry, and sad) by setting α as 30/60/90, setting β as 30/60/90, setting γ as 0, setting ϵ as 30/60/90, and setting λ as 100, resulting in the mixed emotions of delight (surprise+happy), outrage (surprise+angry), and disappointment (surprise+sad), respectively. These mixed emotions are considered more easily perceivable by listeners as indicated in psychological studies [22], [24], [32]. To compare with the mix baseline [24], we use the same two English speakers (0019: female, 0013: male) from the ESD dataset [40], covering five emotions: neutral, angry, happy, sad, and surprise. Each speaker provides 350 utterances per emotion (1,750 total with about 1.2 hours). We follow the official train/validation/test splits [40] and the mix baseline setup [24], with 30 and 20 utterances per emotion in the validation and test sets, respectively. As the dataset contains only single-emotion labels, we set the corresponding emotion parameter (e.g., α for happy) to 100 and others to 0.

Subjective Evaluations: we conduct comprehensive subjective evaluations with 18 listeners to compare our proposed *PUE* with baseline systems. **(1) AB Preference Tests**: listeners select the preferred sample between two systems (A vs. B) based on emotional speech quality. Two tests are conducted: cosy vs. *PUE* and mix vs. *PUE*, each with 24 balanced emotion samples, which cover both single emotion samples and mixed-emotion samples. In mixed samples, surprise is combined with varying proportions of angry, sad, and happy emotions (e.g., surprise + 30% angry, surprise + 60% angry, and surprise + 90% angry). **(2) Best-worst Scaling (BWS) Tests**: listeners identify the most and least preferred samples in sets mixing 100% surprise with varying levels of anger (0%, 30%, 60%, 90%), based on perceived emotional style. **(3) Mean Opinion Score (MOS) Tests**: listeners rate overall speech quality (1 bad–5 excellent) for mixed-emotion expressions (outrage, disappointment, delight) generated by *PUE*, cosy, and mix. MOS tests also evaluate *PUE* for surprise + 0/30/60/90% of other emotions.

TABLE II
BWS RESULT COMPARISON OF THE PROPOSED *PUE* APPROACH ON MIXING 100 % SURPRISE WITH ANGER AT VARYING LEVELS OF ANGRY.

Outrage: perception of angry	**Best Outrage%**	**Worst Outrage%**
Mix surprise with 0% angry	5.56	**80.56**
Mix surprise with 30% angry	0	2.78
Mix surprise with 60% angry	0	11.11
Mix surprise with 90% angry	**94.44**	5.56

Objective Evaluations: we use Whisper-Large-v3 [41] to assess intelligibility by computing WER across: (1) single-emotion test data; (2) surprise mixed with 30%, 60%, and 90% of angry, happy, or sad emotions.

IV. RESULTS AND DISCUSSION

We evaluate *PUE* on emotional speech intelligibility, single and mixed emotion rendering, emotion-guided prompt learning, and zero-shot synthesis of unseen emotional speech.

A. Speech Intelligibility and Single Emotion Generation

Objectively, *PUE* achieves the highest intelligibility across all emotions in the test set (Table I), highlighting *PUE*'s effectiveness in producing clearer emotional speech while preserving the linguistic content. Subjectively, Fig. 2 shows *PUE* surpasses both mix and cosy baselines in perceived quality, confirming its effectiveness in modeling high-quality single-emotion speech.

B. Mixed Emotion Rendering

1) Objective Evaluation: PUE consistently achieves lower WERs across all mixed-emotion cases (Table I), demonstrating its ability to generate more intelligible mixed-emotion speech. The WER values remain more than 20% because mixed-emotion generation, particularly for unseen emotional combinations, is inherently difficult and pushes the limits of current TTS models.

2) Subjective Evaluation: BWS results (Table II) show that 94.44% of listeners identified the mixed-emotion speech containing 90% anger as expressing the strongest sense of outrage compared to versions with lower anger proportions (30% and 60%). 80.56% of participants perceived the nonmixed anger (0%) utterance as the weakest expression of outrage. These demonstrate the controllability of mixed emotion rendering through proportion adjustment. Minor listener disagreement

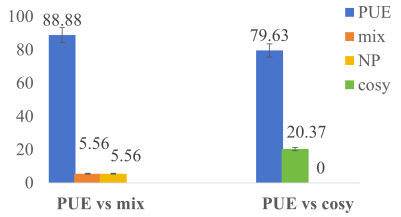

Fig. 2. AB preference test results with 95% confidence interval comparing *PUE* with mix and cosy on seen single emotional speech samples. NP: no preference.

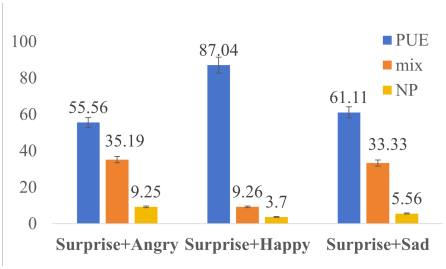

Fig. 3. AB preference test results with 95% confidence interval comparing mix vs. our *PUE* with 100% surprise mixed in varying proportions: 30%, 60%, and 90% for angry, sad, and happy. NP: no preference.

TABLE III
MOS RESULTS COMPARISON WITH 95% CONFIDENCE INTERVAL OF OUR *PUE* UNDER DIFFERENT TEST CASES: ORIGINAL TEST DATABASE; 100% SURPRISE MIXED WITH VARYING PROPORTIONS OF OTHER EMOTIONS: 30%, 60%, AND 90% FOR ANGRY, SAD, OR HAPPY.

Prompt unseen emotions	**MOS**
Mix surprise with 30% angry/sad/happy	3.12±0.16
Mix surprise with 60% angry/sad/happy	2.50±0.13
Mix surprise with 90% angry/sad/happy	**3.57±0.18**

(5.56%) reflects subjective variability likely due to inter-listener variability or the inherent subjectivity of emotional perception. MOS tests (Table III) reveal that the perceptual quality of mixed-emotion speech varies with the proportion of the secondary emotion. Mixing surprise with 30% of another emotion yields moderate quality (3.12), while 60% results in lower naturalness (2.50), likely due to perceptual ambiguity. The best quality is achieved at 90% (3.57), where the secondary emotion becomes more salient and easier to recognize. MOS tests (Table III) further show the best quality when surprise is combined with 90% of another emotion, outperforming lower mixing ratios. This further supports that increasing the proportion of secondary emotions enhances the perception of the target emotion, demonstrating the effectiveness of *PUE* model nuanced emotional blends via flexible parameter settings.

C. Effectiveness of Emotion-guided Prompt Learning

To evaluate the speech quality of the generated mixed-emotion speech, we further conduct AB preference tests comparing our *PUE* against baselines (Fig. 3 and Fig. 4).

1) PUE vs. mix: Fig. 3 shows listener preferences comparing *PUE* and the mix baseline across combinations of surprise with angry, sad, and happy at 30%, 60%, and 90% mix ratios. *PUE* is consistently preferred across all tested configurations, with the highest preference (87.04%) for surprise + happy. This suggests the effectiveness of the proposed emotion-guided prompt learning mechanism in generating perceptually superior mixed emotional speech, especially for acoustically compatible emotions like surprise and happy, which share prosodic traits such as high pitch and energy.

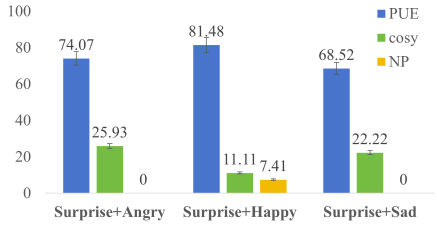

Fig. 4. AB preference test results with 95% confidence interval comparing cosy vs. our *PUE* with 100% surprise mixed in varying proportions: 30%, 60%, and 90% for angry, sad, and happy. NP: no preference.

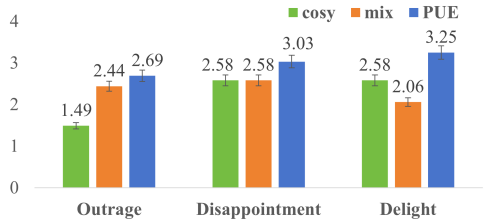

Fig. 5. Comparison of subjective evaluation results for MOS tests with 95% confidence interval across cosy, mix, and the proposed *PUE* models on three unseen emotions: outrage (surprise + 90% angry), disappointment (surprise + 90% sad), and delight (surprise + 90% happy).

2) PUE vs. cosy: *PUE* also outperforms the cosy baseline across all three emotion blends (Fig. 4). For surprise + happy, *PUE* again achieved the highest preference 81.48% versus 11.11% for cosy, reinforcing earlier findings from comparisons with the mix baseline. These consistent results confirm the effectiveness of prompt-based emotion conditioning in learning the interplay between different emotions, yielding speech that is more natural, emotionally rich, and preferred by human listeners over existing baseline systems.

D. Zero-shot Unseen Emotional Speech Synthesis

We further assess the zero-shot ability of *PUE* using three unseen composite emotions: outrage (surprise + 90% angry), disappointment (surprise + 90% sad), and delight (surprise + 90% happy). Fig. 5 reports MOS scores, comparing *PUE* with cosy and mix baselines. *PUE* achieves the highest MOS across all cases, underscoring its zero-shot generalization strength. The superior performance reflects the compositional strength of emotion-guided prompts, which allow *PUE* to model unseen emotions by leveraging known emotion components and synthesize novel emotional states in a controllable manner. The best MOS of 3.25 highlights the inherent challenge of modeling unseen emotional speech, which we will address in future work through potential self-training and data augmentation strategies. Synthesized samples are available at the link[1].

V. CONCLUSION

This paper presents an innovative *PUE* approach for expressive speech synthesis through emotion-guided prompt learning on an LLM-TTS architecture. The proposed *PUE* significantly advances expressive TTS systems by enabling the generation of unseen emotional speech through modifying categorical emotion-relevant prompts, eliminating the need for additional training on new emotion categories. Both subjective and objective evaluations demonstrate that *PUE* outperforms state-of-the-art baselines, demonstrating its effectiveness in capturing diverse emotional nuances and facilitating zero-shot emotion generation.

[1]https://xiaoxue1117.github.io/PUE/

REFERENCES

[1] Y. Yasuda and T. Toda, "Text-to-speech synthesis based on latent variable conversion using diffusion probabilistic model and variational autoencoder," in *IEEE ICASSP*, 2023, pp. 1–5.

[2] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 644–12 652.

[3] F. Khanam, F. A. Munmun, N. A. Ritu, A. K. Saha, and M. Firoz, "Text to speech synthesis: A systematic review, deep learning based architecture and future research direction," *Journal of Advances in Information Technology Vol*, vol. 13, no. 5, 2022.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.

[5] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[6] Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," *arXiv preprint arXiv:2410.17196*, 2024.

[7] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *IEEE ICASSP*, 2020, pp. 7254–7258.

[8] D. Diatlova and V. Shutov, "Emospeech: guiding fastspeech2 towards emotional text to speech," in *SSW 2023*.

[9] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.

[10] X. Li, Z.-Q. Cheng, J.-Y. He, X. Peng, and A. G. Hauptmann, "Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis," *arXiv preprint arXiv:2404.18398*, 2024.

[11] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *IEEE ICASSP*, 2023, pp. 1–5.

[12] 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 Transactions on Audio, Speech, and Language Processing*, 2024.

[13] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[14] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *IEEE ICASSP*, 2021, pp. 5734–5738.

[15] H. Wu, X. Wang, S. E. Eskimez, M. Thakker, D. Tompkins, C.-H. Tsai, C. Li, Z. Xiao, S. Zhao, J. Li *et al.*, "Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech," *arXiv preprint arXiv:2407.12229*, 2024.

[16] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive text-to-speech using style tag," *arXiv preprint arXiv:2104.00436*, 2021.

[17] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *ISCSLP*. IEEE, 2021, pp. 1–5.

[18] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[19] X. Gao, Y. Chen, X. Yue, Y. Tsao, and N. F. Chen, "Ttslow: Slow down text-to-speech with efficiency robustness evaluations," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[20] S. Inoue, K. Zhou, S. Wang, and H. Li, "Hierarchical control of emotion rendering in speech synthesis," *arXiv preprint arXiv:2412.12498*, 2024.

[21] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[22] A. Braniecka, E. Trzebińska, A. Dowgiert, and A. Wytykowska, "Mixed emotions and coping: The benefits of secondary emotions," *PloS one*, vol. 9, no. 8, p. e103940, 2014.

[23] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Emomix: Emotion mixing via diffusion models for emotional speech synthesis," *arXiv preprint arXiv:2306.00648*, 2023.

[24] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2022.

[25] Y. Wang, D. Stanton, Y. Zhang, R.-S. 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 *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[26] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *2021 IEEE SLT*, 2021, pp. 423–430.

[27] R. Liu, Y. Hu, Y. Ren, X. Yin, and H. Li, "Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 698–18 706.

[28] W. Zhao and Z. Yang, "An emotion speech synthesis method based on vits," *Applied Sciences*, vol. 13, no. 4, p. 2225, 2023.

[29] K. Zhou, Y. Zhang, S. Zhao, H. Wang, Z. Pan, D. Ng, C. Zhang, C. Ni, Y. Ma, T. H. Nguyen *et al.*, "Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions," *arXiv preprint arXiv:2409.16681*, 2024.

[30] X. Gao, C. Zhang, Y. Chen, H. Zhang, and N. F. Chen, "Emo-dpo: Controllable emotional speech synthesis through direct preference optimization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[31] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. DAI *et al.*, "Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis," *arXiv preprint arXiv:2502.04128*, 2025.

[32] R. Plutchik and H. Kellerman, *Theories of emotion*. Academic press, 2013, vol. 1.

[33] M. Cross and C. Hanrahan, *Changing Minds: The Go-to Guide to Mental Health for Family and Friends*. HarperCollins Australia, 2016.

[34] K. Zhou, B. Sisman, C. Busso, B. Ma, and H. Li, "Mixed-evc: Mixed emotion synthesis and control in voice conversion," *arXiv preprint arXiv:2210.13756*, 2022.

[35] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[36] X. Gao, H. Zhang, and N. F. Chen, "Multigen: Child-friendly multilingual speech generator with llms," *arXiv preprint arXiv:2508.08715*, 2025.

[37] G. T. Google, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[38] Y. Chen, X. Yue, X. Gao, C. Zhang, L. F. D'Haro, R. Tan, and H. Li, "Beyond single-audio: Advancing multi-audio processing in audio large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 10 917–10 930.

[39] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[40] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[41] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356