

Word-level Emotional Intensity Control in TTS via Emotion Residual Vectors

Ji-Hyun Park ¹, Nam-Seok Song ², Joon-Hyuk Chang ^{1,2,**}

¹ Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea

² Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea

{jjhh0210, insang99, jchang}@hanyang.ac.kr

Abstract

Modern emotional text-to-speech (TTS) models typically rely on utterance-level conditioning. However, natural emotional speech requires fine-grained variations across words to achieve nuanced, human-like expressiveness. To address this, we introduce emotion residual vectors (ERVs), which capture neutral-to-emotional deviations in word-aligned self-supervised speech embeddings as an annotation-free cue for local prosody. To stabilize intensity control, we compress ERVs into a low-dimensional bottleneck space and train a RoBERTa-based predictor to infer the projected vectors from text and emotion. During inference, the predicted vectors are converted into additive hidden-state offsets and injected into a neutral-conditioned TTS model, enabling continuous word-level intensity control. Experimental results show that our method improves local emotional controllability while preserving naturalness compared to controllable emotional TTS baselines.

Index Terms: Emotional text-to-speech, emotion intensity, word-level emotion control

1. Introduction

Emotional expressiveness is important for the perceived naturalness of text-to-speech (TTS). In prior work on emotional TTS [1–6], utterance-level emotion conditioning was commonly adopted due to its implementation simplicity and straightforward controllability. However, when the conditioning is applied only at the utterance level, the model can misallocate word-level prosodic prominence, emphasizing non-target words while attenuating the intended focus word. Consequently, subsequent studies have investigated fine-grained conditioning schemes [7–9] that parameterize emotional intensity as a scalar control variable and apply it through word-aligned conditioning, thereby allowing per-word adjustment of emotion strength.

Representative methods include EmoQ-TTS [7], which enables local intensity control by deriving proxy intensity levels from the embedding-space distance to a neutral reference and conditioning on quantized codes during phoneme-aligned synthesis. HED-TTS [8] further extends controllability by predicting multi-granular emotion distributions and allowing direct manipulation at the utterance, word, or phoneme level, using ranking-based scoring to derive normalized intensity targets for supervision. EME-TTS [9] additionally introduces explicit emphasis control and mitigates emotion–emphasis interference by reweighting attention within a masked emphasis region. While these designs enable precise control, they can also degrade perceptual quality, making the synthesized speech sound less natural. In particular, word-level control may lead to local prosodic

distortions, including abrupt pitch shifts, unnatural durations, and overly salient target words. Consequently, maintaining perceptual naturalness remains challenging for existing word-level emotion control methods.

To address this limitation, we model the neutral-to-emotional deviation itself as a word-level control signal. We define emotion residual vectors (ERVs) as word-aligned residuals in a pretrained self-supervised speech (S3L) representation space. We adopt S3L representations because they provide rich paralinguistic and prosodic cues without requiring manual prosody annotations [10–14]. Recent work [15] shows that S3L residuals between matched word instances encode structured prosodic shifts and are concentrated in a compact subspace. Based on this observation, we interpret the difference between aligned emotional and neutral words as a local signal of emotional prosody. We inject ERVs into a neutral-conditioned TTS backbone and control intensity by scaling the injected deviation with word-level control weights. Based on this formulation, our work makes the following three contributions:

1. We show that ERV based injection enables word-level emotional intensity control while preserving naturalness.
2. We propose a bottleneck residual subspace that provides a more stable modulation signal for intensity scaling.
3. We propose a three-stage training pipeline that integrates ERV extraction, bottleneck injection learning, and text-and-emotion based prediction for controllable emotional TTS.

Audio samples are available online.¹

2. Proposed Method

In this section, we describe our overall procedure for deriving ERVs and integrating them into a TTS system. The overall architecture is illustrated in Fig. 1(a). The proposed framework follows a three-stage training scheme:

1. Pretrain the FastSpeech2 [16] backbone with speaker and utterance-level emotion conditioning.
2. Freeze the backbone and train a residual injection module (RIM) with the global emotion fixed to neutral.
3. Construct projected-ERV targets using the fixed bottleneck projection and fine-tune a RoBERTa-based [17] predictor to estimate word-level projected ERVs.

During inference, the predictor outputs projected ERVs conditioned on text and the target emotion. These vectors are injected into the neutral-conditioned backbone and scaled by control weights, enabling word-level emotional intensity control.

^{**} indicates the corresponding author.

¹<https://jjhh0210.github.io/EmoRes-tts-demo/>

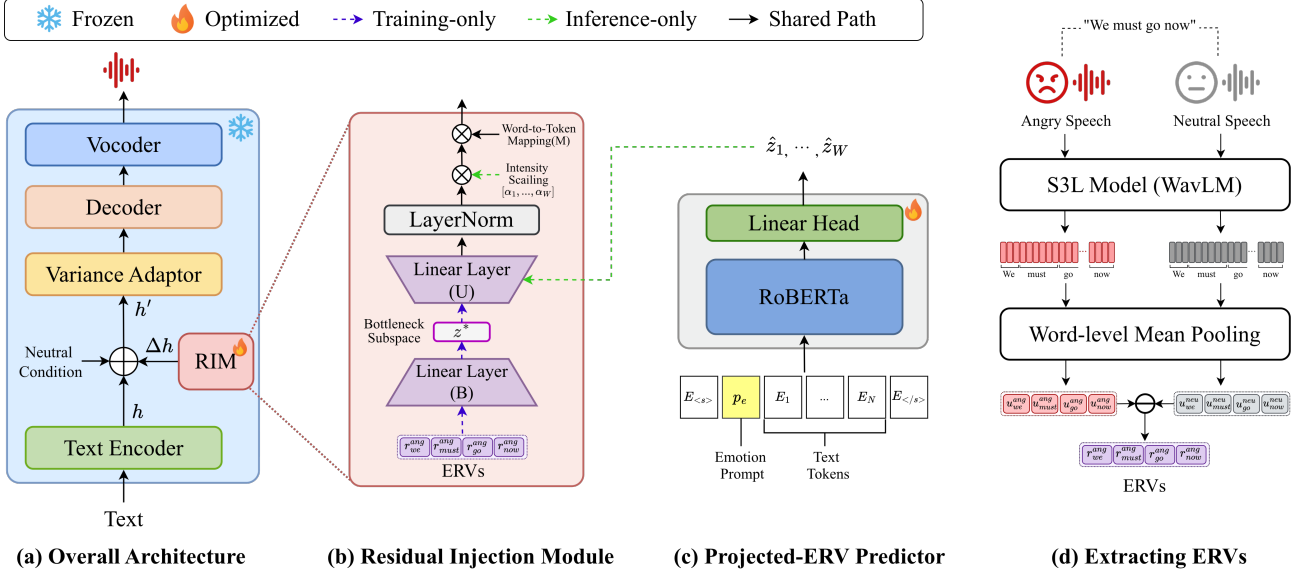


Figure 1: Overview of the proposed ERV-based emotional TTS framework. (a) Overall architecture with a residual injection module (RIM). (b) Structure of the RIM. (c) RoBERTa-based projected-ERV predictor. (d) ERVs extraction process.

2.1. Emotion Residual Vectors (ERVs)

We represent local neutral-to-emotional deviations using S3L embeddings, which provide robust acoustic representations without requiring prosody annotations. Based on word-aligned embeddings, we define ERVs as residual vectors between matched neutral and emotional word instances. The ERV extraction procedure is illustrated in Fig. 1(d). For each speaker and transcript in emotional speech database (ESD) [18], we form neutral–emotional pairs by matching utterances that share identical text content but differ in emotion label. Word boundaries are obtained via montreal forced aligner (MFA) [19]. Then we extract frame-level embeddings from a pretrained S3L model [20]. Following a residual-space approach that compares paired neutral and emphasized word instances [15], we compute a word-level embedding for each word w by mean pooling the frame embeddings within its aligned interval, yielding $\mathbf{u}_w^{\text{neu}}$ and $\mathbf{u}_w^{\text{emo}}$. We define the ERV as

$$\mathbf{r}_w = \mathbf{u}_w^{\text{emo}} - \mathbf{u}_w^{\text{neu}}. \quad (1)$$

We interpret this residual as a local representation of the shift from neutral to emotional speech in S3L space.

2.2. Residual Injection Module

We incorporate a RIM into a pretrained FastSpeech2 [16] backbone to convert word-level ERVs into additive modulation signals in the encoder hidden space. The module structure is shown in Fig. 1(b). The TTS text encoder produces hidden states $\mathbf{h}_{1:T}$ with $\mathbf{h}_t \in \mathbb{R}^C$, where T denotes the token sequence length and C the hidden dimension.

Each word-level ERV $\mathbf{r}_w \in \mathbb{R}^D$ is transformed into an injection vector in the encoder hidden space through bottleneck projections:

$$\Delta \mathbf{h}^{(w)} = \text{LayerNorm}(\mathbf{U} \mathbf{B} \mathbf{r}_w), \quad (2)$$

where $\mathbf{B} \in \mathbb{R}^{a \times D}$ projects the high-dimensional residual into a low-dimensional bottleneck subspace ($a < D$), and $\mathbf{U} \in \mathbb{R}^{C \times a}$

maps the bottleneck representation to the encoder hidden dimension. Layer normalization [21] stabilizes modulation magnitude, making intensity scaling more consistent.

Word-level intensity is controlled by a nonnegative vector $\alpha = [\alpha_1, \dots, \alpha_N]$ and distributed to token positions using a word-to-token mapping matrix $\mathbf{M} \in \{0, 1\}^{T \times N}$:

$$\Delta \mathbf{h}_t = \sum_{w=1}^N M_{t,w} \alpha_w \Delta \mathbf{h}^{(w)}. \quad (3)$$

The modulated encoder state at the token position t is then formed as

$$\mathbf{h}'_t = \mathbf{h}_t + \mathbf{c}_{\text{neu}} + \Delta \mathbf{h}_t, \quad (4)$$

where $\mathbf{c}_{\text{neu}}$ is constructed by projected concatenation of the speaker embedding and the neutral emotion embedding.

To learn the bottleneck injection space, we freeze the FastSpeech2 [16] backbone and fix the global emotion condition to neutral. We then optimize only the RIM parameters $\{\mathbf{B}, \mathbf{U}\}$ (and LayerNorm parameters) with the standard FastSpeech2 reconstruction loss to match the corresponding emotional target speech.

2.3. Projected-ERV Predictor

Using the learned bottleneck projection $\mathbf{B}$, we predict ERVs in the bottleneck space instead of regressing high-dimensional residuals. Fig. 1(c) shows the predictor architecture. This design provides a compact and stable control signal that interfaces directly with the residual injection module.

To supervise the predictor, we first construct word-level targets by projecting each ERV through the learned bottleneck projection:

$$\mathbf{z}_w^* = \mathbf{B} \mathbf{r}_w, \quad (5)$$

where $\mathbf{z}_w^* \in \mathbb{R}^a$ denotes the projected ERV for word w . We then train a RoBERTa-based [17] predictor to estimate $\hat{\mathbf{z}}_w$ from text and a target emotion e , leveraging its bidirectional contextual representations to model word-level variation conditioned

on emotion. We form the RoBERTa input sequence by inserting a trainable emotion prompt embedding $\mathbf{p}_e$ after the start token:

$$[\langle s \rangle, \mathbf{p}_e, \mathbf{x}, \langle /s \rangle], \quad (6)$$

where $\mathbf{p}_e \in \mathbb{R}^{d_{\text{nb}}}$ with d_{nb} the RoBERTa hidden size, and $\mathbf{x} = (x_1, \dots, x_N)$ is the BPE-based subword token sequence.

RoBERTa produces contextualized representations at the subword level. We obtain word-level features by mean pooling the hidden states of subword tokens aligned to each word, and use a linear regression head to map the pooled vectors to the projected ERV for each word. All RoBERTa layers are fine-tuned jointly with the emotion prompt embedding and the linear regression head. The predictor is optimized using a word-wise mean squared error objective:

$$\mathcal{L}_z = \frac{1}{W} \sum_{w=1}^W \|\hat{\mathbf{z}}_w - \mathbf{z}_w^*\|_2^2, \quad (7)$$

where W is the number of words in the utterance, $\hat{\mathbf{z}}_w$ is the predicted projected ERV for word w , and $\mathbf{z}_w^*$ is the corresponding target projection defined in Eq. (5).

2.4. Intensity Control at Inference

At inference, the predictor outputs word-level projected ERVs $\hat{\mathbf{z}}_w$, which are mapped to hidden-state modulations via the frozen injection module and injected into the neutral-conditioned backbone ($\Delta \mathbf{h}^{(w)} = \text{LayerNorm}(\mathbf{U}\hat{\mathbf{z}}_w)$). Word-level intensity is controlled through the scaling vector α in Eq. (3), while the global emotion condition is fixed to neutral.

3. Experiments and Results

3.1. Data

We conducted experiments on the English subset of emotional speech database (ESD) [18], which contains 10 speakers and five emotions (neutral, angry, happy, sad and surprise). ESD provides parallel utterances with identical transcripts across emotions, enabling same-speaker neutral-emotional pairing. We followed the train/validation/test split used in prior work [9] on ESD (14,900 / 950 / 1,550 utterances). Using these parallel pairs, we extracted 88,400 word-level ERV instances, 22,100 per non-neutral emotion.

3.2. Implementation Details

The acoustic backbone followed FastSpeech2 [16] with four FFT blocks in both encoder and decoder, hidden size 256, two attention heads, filter size 1024, and convolution kernel sizes (9,1). We trained the model with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$, and weight decay = 0) for 200,000 steps with a batch size of 64. We employed an iSTFTNet vocoder [22] trained on ESD with 16 kHz mel-spectrograms. It adopts the C8C8I architecture with two upsampling layers of factors 6 and 8. Mel-spectrograms were extracted by applying the short-time Fourier transform (STFT) to 16 kHz audio with hop length 192, window length 768, FFT size 768, and an 80-bin Mel filterbank. To extract ERVs, we used WavLM-Base [20] features and aggregated layers 7–12 to form 768-dimensional word-level residual targets. We set the bottleneck dimension to $a = 32$ by default, and varied it in our ablation studies. For the projected-ERV predictor, we fine-tuned a pretrained RoBERTa-Base [17] encoder using AdamW for 20 epochs with a batch size of 8. We set the learning rate to $\eta_{\text{RoBERTa}} = 5 \times 10^{-5}$ for the RoBERTa parameters and $\eta_{\text{head}} = 3 \times 10^{-4}$ for the prediction head.

Table 1: Subjective and objective evaluation results on naturalness and emotion expressiveness of synthesized speech.

Model	Subjective		Objective	
	NMOS $\uparrow$	EMOS $\uparrow$	UTMOS $\uparrow$	Emo.Acc. $\uparrow$
GT	4.73 $\pm$ 0.15	4.04 $\pm$ 0.19	4.17 $\pm$ 0.03	0.72
iSTFTNet [22]	4.56 $\pm$ 0.17	4.21 $\pm$ 0.17	4.11 $\pm$ 0.03	0.74
FS2+emo [16]	3.81 $\pm$ 0.19	3.61 $\pm$ 0.21	3.70 $\pm$ 0.04	0.70
EME-TTS [9]	3.78 $\pm$ 0.19	3.50 $\pm$ 0.20	3.76 $\pm$ 0.03	0.71
HED-TTS [8]	2.75 $\pm$ 0.19	2.97 $\pm$ 0.21	2.90 $\pm$ 0.04	0.54
Proposed	3.83 $\pm$ 0.19	3.56 $\pm$ 0.21	3.77 $\pm$ 0.04	0.71

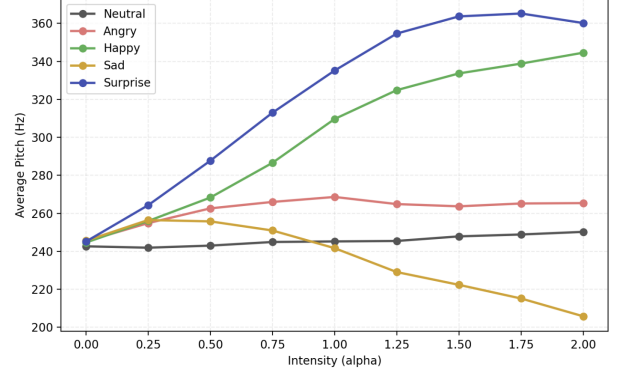


Figure 2: Average pitch tendency for each emotion as the intensity factor α increases.

3.3. Model Performance

We conducted both subjective and objective evaluations for overall speech quality and emotional expression. For subjective evaluation, 15 listeners rated naturalness with the Naturalness Mean Opinion Score (NMOS) and emotional expressiveness with the Emotional Mean Opinion Score (EMOS) on a 5-point scale. For objective evaluation, we computed the UTokyo Mean Opinion Score predictor (UTMOS) [23] and emotion classification accuracy (Emo.Acc.) with a pretrained Emotion2vec-plus-large model [24] on synthesized samples from the ESD test set.

We compared the proposed method with FastSpeech2-based emotion-controllable baselines: (i) FS2+emo, our FastSpeech2 backbone trained with standard utterance-level emotion embeddings; (ii) EME-TTS, which supports emotion control with word-level emphasis conditioning; and (iii) HED-TTS, which predicts hierarchical emotion distributions for word and phoneme-level control. For a fair comparison of utterance-level emotional rendering, we set the word-level control weights to $\alpha_w = 1$ for all words in our model. Table 1 shows that the proposed method achieved comparable performance to the FS2+emo backbone, while substantially outperforming baselines. These results indicate that ERV-based modulation preserved overall speech quality and emotional expressiveness.

We further analyzed the effect of the intensity scaling factor α on prosody. For utterance-level intensity control, Figure 2 shows that as α increases, the average pitch changes in emotion-specific directions (e.g., increasing for happy/surprise and decreasing for sad). Although the model was trained with $\alpha = 1$ on all emotion categories, we observed that the pitch trends remain directionally consistent even when $\alpha > 1$ at inference.

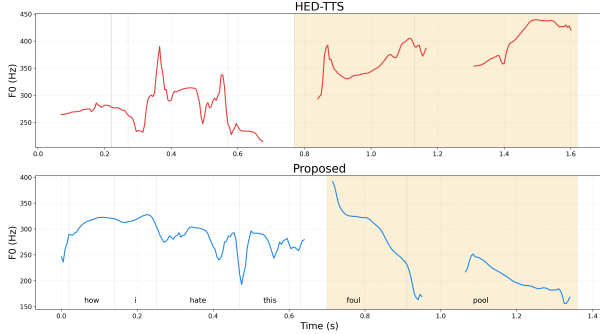


Figure 3: F_0 contours for an utterance synthesized by HED-TTS and the proposed method under the same target emotion. The last two words (“foul”, “pool”) are highlighted.

Table 2: A/B preference results for naturalness of highlighted-word emphasis within an utterance (Win–Tie–Lose).

Model	Win–Tie–Lose
Proposed vs EME-TTS [9]	55.3% – 16.7% – 28.0%
Proposed vs HED-TTS [8]	87.1% – 7.5% – 5.4%

3.4. Word-level Intensity Controllability

We conducted an A/B preference test with 15 listeners to evaluate word-level controllability and its naturalness. Given paired samples from our model and word-level controllable baselines, listeners selected which sample renders the highlighted words more naturally within the sentence. We reported the results in a Win–Tie–Lose format, where *Win* is the percentage of trials preferring our method, *Lose* prefers the baseline, and *Tie* indicates no preference. As shown in Table 2, the proposed method was preferred over the baselines, indicating improved perceptual naturalness for localized intensity control.

To further analyze the preference results, we compared prosodic realizations of a representative example. For this analysis, we assigned higher control weights to target words (e.g., $\alpha_w = 1.0$) and set the other words to a lower value (e.g., $\alpha_w = 0.5$). Fig. 3 visualizes the pitch contour of an example utterance in the *happy* condition, where the target word region is highlighted in yellow. The proposed method increased pitch around the target words while maintaining smooth transitions with surrounding regions, resulting in a more natural emphasis intonation within the utterance. In contrast, HED-TTS produced a local pitch excursion at the target words, which reduced prosodic coherence with the surrounding context in this example.

3.5. Ablation studies

To examine the effect of the bottleneck subspace and dimension a , we conducted ablation studies with three settings.

No bottleneck. With the backbone frozen, we removed the bottleneck and trained only a single projection that maps the ground-truth ERVs directly to the encoder hidden space.

Bottleneck dimension sweep. We trained the injection module parameters (B , U , and LayerNorm) while varying the bottleneck dimension $a \in \{4, 16, 32, 64, 128\}$. Each setting injected test-set ERVs after projection to z .

Proposed. After selecting $a = 32$, we trained a predictor to

Table 3: Ablation studies on bottleneck design and predictor.

Method	WER ↓	UTMOS ↑	Emo.Acc. ↑
No bottleneck	7.53	3.71±0.04	0.55
$a = 4$	9.00	3.75±0.04	0.47
$a = 16$	7.97	3.74±0.04	0.57
$a = 32$	8.22	3.71±0.04	0.62
$a = 64$	8.29	3.73±0.03	0.61
$a = 128$	8.24	3.73±0.03	0.61
Proposed (w/ predictor)	8.03	3.75±0.04	0.71

estimate z from text and emotion, and evaluated the full system by injecting the predicted z into the acoustic model.

Table 3 summarizes the ablation results. We computed word error rate (WER) using the Whisper ASR model [25] with the faster-whisper implementation. Overall, the UTMOS varied slightly across settings (3.71–3.75), and most differences fall within the reported confidence intervals. This indicates that the bottleneck design mainly affected emotion rendering rather than overall speech quality. Comparing *no bottleneck* to bottlenecked injection, directly projecting the ground-truth ERV to the encoder space had the lowest WER, while Emo.Acc. increased when ERVs were injected through a bottleneck with moderate capacity (e.g., $0.55 \rightarrow 0.62$ at $a = 32$). This suggests that high-dimensional residual injection is not necessary to obtain competitive quality, and that a compact bottleneck representation can provide a more effective control signal for emotion rendering. In the bottleneck dimension sweep, $a = 4$ had the lowest Emo.Acc. (0.47), while $a \geq 32$ saturated (0.61–0.62). Further increasing the bottleneck beyond $a = 64$ resulted in little change in Emo.Acc. (0.61 at $a = 64$ and 0.61 at $a = 128$), suggesting that additional bottleneck capacity brought limited benefit. Thus, we used $a = 32$ by default. Finally, the predictor-based system achieved the highest Emo.Acc. (0.71) with comparable UTMOS. This was expected because the predictor generates z conditioned on the target emotion, whereas injection-only settings rely on residual modulation with the global emotion fixed to neutral.

4. Conclusion

In this paper, we proposed ERVs and a bottlenecked residual injection framework for word-level emotional intensity control in TTS. Experiments indicated that ERVs captured emotion-related prosodic deviations, and that projecting ERVs into a low-dimensional bottleneck preserves most of the modulation information useful for synthesis. The bottleneck representation also simplified prediction, enabling a RoBERTa-based predictor to estimate projected ERVs and provide word-level intensity control while maintaining perceptual naturalness (e.g., improved A/B preference for highlighted-word naturalness). These findings suggest the feasibility of a practical pipeline in which high-dimensional S3L representations are compressed into a compact space that remains predictable and controllable, rather than being handled directly in the original space.

A limitation of our approach is its reliance on parallel, word-aligned neutral–emotional pairs for ERV extraction. Nonetheless, our results show that such residuals provide an informative and structured signal for prosody control. Extending residual-based control to less constrained data settings is an important direction for future work.

5. Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

6. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and polishing of the manuscript.

7. References

- [1] 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, 2023.
- [2] S. Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H. G. Kang, "Emotional speech synthesis with rich and granularized control," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7254–7258.
- [3] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448–1460, 2022.
- [4] D. Diatlova and V. Shutov, "Emospeech: Guiding fastspeech2 towards emotional text to speech," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW)*, 2023, pp. 106–112.
- [5] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee, "EmoSphere-TTS: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech," in *Proc. Interspeech*, 2024, pp. 1810–1814.
- [6] D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee, "EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 2365–2380, 2025.
- [7] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "EmoQ-TTS: Emotion intensity quantization for fine-grained controllable emotional text-to-speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6317–6321.
- [8] S. Inoue, K. Zhou, S. Wang, and H. Li, "Hierarchical emotion prediction and control in text-to-speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 601–10 605.
- [9] H. Li, L. Qu, J. Hu, and T. Li, "EME-TTS: Unlocking the emphasis and emotion link in speech synthesis," in *Proc. Interspeech*, 2025, pp. 4368–4372.
- [10] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.
- [11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [13] S.-w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.
- [14] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.
- [15] S. R. Cassini, T. Hain, and A. Ragni, "Emphasis sensitivity in speech representations," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2025.
- [16] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations (ICLR)*, 2021.
- [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [18] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kald," in *Proc. Interspeech*, 2017, pp. 498–502.
- [20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Li, N. Kanda, T. Yoshioka, X. Xiao, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [21] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [22] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6207–6211.
- [23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for Voice-MOS challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.
- [24] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 15 747–15 760.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.