APSIPA Transactions on Signal and Information Processing, 2025, 14, e302

Original Paper

Multi-step Prediction and Control of Hierarchical Emotion Distribution in Text-to-speech Synthesis

Sho Inoue[1,3], Kun Zhou[4], Shuai Wang[2] and Haizhou Li[1,3,5*]

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China*
[2] *School of Intelligence Science and Technology, Nanjing University, Suzhou, China*
[3] *Shenzhen Research Institute of Big Data, Shenzhen, China*
[4] *Tongyi Speech Lab, Alibaba Group, Singapore*
[5] *Department of ECE, National University of Singapore, Singapore*

ABSTRACT

We investigate hierarchical emotion distribution (ED) for achieving multi-level quantitative control of emotion rendering in text-to-speech synthesis (TTS). We introduce a novel multi-step hierarchical ED prediction module that quantifies emotion variance at the utterance, word, and phoneme levels. By predicting emotion variance in a multi-step manner, we leverage global emotional context to refine local emotional variations, thereby capturing the intrinsic hierarchical structure of speech emotion. Our approach is validated through its integration into a variance adaptor and an external module design compatible with various TTS systems. Both objective and subjective evaluations demonstrate that the proposed framework significantly enhances emotional expressiveness and enables precise control of emotion rendering across multiple speech granularities.

*Corresponding author: haizhouli@cuhk.edu.cn

Received 28 February 2025; revised 09 June 2025; accepted 03 July 2025
ISSN 2048-7703; DOI 10.1561/116.20250010

Keywords: Hierarchical emotion distribution, multi-step emotion prediction, text-to-speech synthesis

1 Introduction

Text-to-speech (TTS) synthesis focuses on generating human-like speech from text input [38]. Advancements in deep learning have significantly improved the naturalness and quality of synthesized speech. However, current TTS systems still struggle with conveying emotional expressiveness and precisely controlling emotional nuances, limiting their ability to deliver humanlike expressive speech [41]. To address these limitations, Emotional TTS aims to bridge this gap by enhancing speech expressiveness, enabling more engaging and empathetic dialogue systems with emotional intelligence [40].

Emotional TTS faces challenges stemming from the hierarchical structure of human emotions [22]. Since speech emotion is characterized by distinct prosodic patterns at the phoneme, word, and utterance levels [41, 22, 9], these patterns naturally form a hierarchy, as established in previous studies [6, 35]. The prior literature indicates that modifying only global prosodic attributes does not capture the full complexity of emotional speech [48, 46, 23, 49]. Additionally, prior work in text-to-speech synthesis and emotional voice conversion underscores the necessity of multi-level modeling [29, 49, 24]. Consequently, developing a method to model the hierarchical structure of emotions is essential for generating nuanced speech synthesis. However, existing text-based emotion representation prediction networks in controllable models address phoneme-level variations [33, 25], overlooking the benefits of multi-level emotion modeling.

In this work, we build upon our previous work on a multi-level quantifiable method for speech emotion control [14] and editing [12] by proposing a multi-step prediction framework for hierarchical emotion distribution (ED) derived from textual cues. Our proposed pipeline supports three inference scenarios, as illustrated in Figure 1: (a) Text-to-Speech (TTS) with Emotion Prediction, where the hierarchical emotion distribution (ED) is directly predicted from the input text; (b) TTS with Emotion Control, where the ED is predicted from the text and can be modified by users; and (c) Emotion Editing, where the ED is extracted from input audio and manually adjusted by users. In [14, 12, 13], a hierarchical ED was introduced to enable both global and fine-grained emotion modification in speech generation. Unlike prior single-step ED prediction approaches [14], which treat different levels of emotion variance independently, we propose to explicitly model hierarchical dependencies by predicting EDs at the utterance, word, and phoneme levels in a multi-step manner. This structured approach ensures that higher-level emotional context influences lower-level prosodic details, resulting in a more coherent,

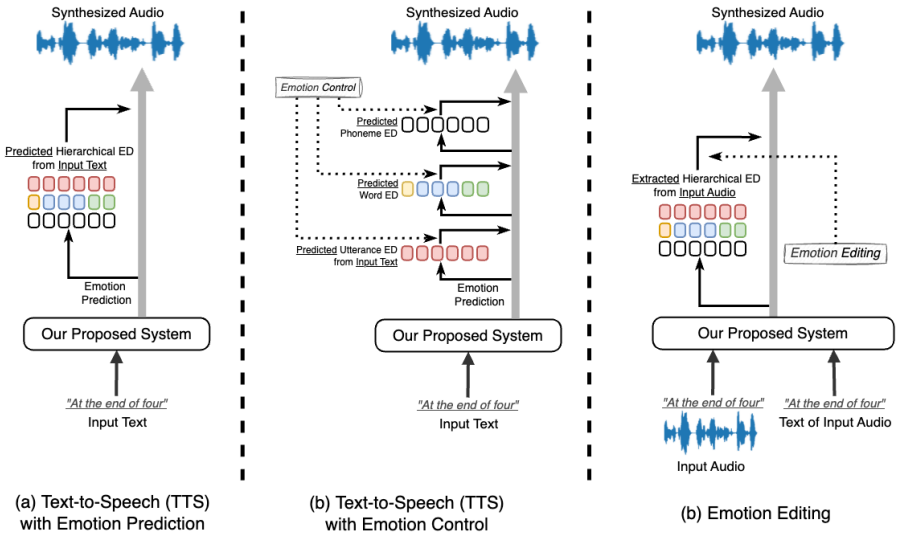

Figure 1: Inference diagram of the proposed system: (a) TTS with emotion prediction; (b) TTS with emotion control; and (c) emotion editing. The hierarchical emotion distribution (ED) can be obtained in three ways: (1) directly predicted from the input text ("Emotion Prediction"), (2) predicted from the input text with user modifications ("Emotion Control"), or (3) extracted from the input audio and manually adjusted by users ("Emotion Editing").

expressive, and controllable emotional rendering. By leveraging multi-step ED prediction, our method provides fine-grained control, closely mimicking the way humans modulate speechstarting with an overall tone and refining intonation and articulation dynamically. This leads to an improved performance on both emotion expressiveness and speech naturalness. Furthermore, to demonstrate its flexibility, we explore two integration strategies for hierarchical ED: implementing it as a variance adaptor within FastSpeech2 [14] and incorporating it as an external module compatible with any text-to-speech (TTS) model [12]. Through this approach, we bridge the gap between interpretability and fine-grained emotion control. Our contributions are summarized as follows:[1]

- We introduce a multi-step prediction framework for hierarchical emotion distribution (ED), where the utterance-, word-, and phoneme-level EDs are derived from textual cues in successive steps. This structured approach ensures that higher-level emotional context guides low-level prosodic details, resulting in synthesized emotional speech that is both natural and expressive;

[1]**Demo**: https://shinshoji01.github.io/multi-step-prediction-HED/.

- Leveraging the multi-step prediction of EDs, our method achieves refined control over emotion rendering and closely emulates human speech modulation. This approach not only enhances global and nuanced emotional expressiveness but also significantly improves performance in controlled emotional voices;
- We explore two integration strategies for hierarchical ED into TTS systems: (1) embedding it as a variance adaptor within FastSpeech2 and (2) implementing it as an external module, making it adaptable to various text-to-speech (TTS) systems.

The rest of this paper is organized as follows: In Section 2, we discuss the related works. Section 3 describes our proposed methodology. In Section 4, we summarize experimental setups. In Section 5, we report our experiments and results. Section 6 concludes our study.

2 Related Works

In this section, we briefly introduce related studies to set the stage for our research and highlight the novelty of our contributions. We begin by discussing the hierarchical nature of speech emotions, emphasizing the need for multi-level and multi-step emotion modeling. We then review existing approaches to emotion rendering control in the TTS literature, identifying key advancements and gaps that our work addresses.

2.1 Hierarchical Nature of Speech Emotion

Speech emotions manifest hierarchically across three levels: utterance, word, and phoneme. At the utterance level, previous studies have shown that global prosodic patterns—such as pitch contour, range, mean, intonation, tempo, and rhythm—play a crucial role in conveying emotion [34]. At the word level, lexical cues shape emotional tone [45] and enhance intensity through emphasis [10]. Additionally, research suggests that when lexical and prosodic signals conflict, listeners tend to rely on prosodic cues to interpret emotion [37]. At the phoneme level, individual prosodic features such as pitch, energy, and duration contribute to emotional expression [20], as supported by multiple studies [1, 7, 26]. This hierarchical structure highlights the necessity of studying multi-level and multi-step emotion modeling for effective emotion modeling and control.

2.2 Control of Speech Emotion

Recent advancements in emotional TTS have significantly enhanced expressiveness; however, achieving interpretable emotion control remains a challenge.

Prior studies have primarily focused on refining emotion intensity control by treating speech emotion as a global feature and manipulating representations or attributes derived from reference audio. For instance, [30] enables utterance-level control via a speech mixer that predicts pseudo-labels and modulates features such as pitch, duration, and energy. [47, 27] further controlled emotion by manipulating speaker-disentangled representations in cross-speaker scenarios. Recent studies employ relative attributes [31] for intensity control [53, 55]. Approaches for mixed emotions include manipulating relative attributes [52], incorporating noise mixing in diffusion models [39], and leveraging continuous emotional representations [54]. In contrast, EmoSphere-TTS [2, 3] models emotional complexity via a spherical emotion vector through Cartesian-spherical transformations, while [15] integrates computational paralinguistic text prompts to enhance emotional expressiveness.

To achieve fine-grained emotion control, several studies have explored segmental-level representations. For example, MsEmoTTS [25] employs a global emotion label and modifies phoneme-level intensity via relative attributes. EmoQ-TTS [11] quantifies emotion intensity using a distance-based method, while the study in [44] refines control by examining inter- and intra-class distances. Additionally, CASEIN [5] leverages a speech emotion recognition module to predict phoneme-level emotion distributions. These multi or phonemelevel approaches outperform utterancelevel modeling in emotional speech synthesis [25, 11, 5] Building upon these efforts, our previous work introduced hierarchical emotion distribution (ED)[14, 12], which enables multi-level emotion control, capturing both global and fine-grained emotional variations in speech synthesis.

However, existing approaches to hierarchical emotion modeling still face several limitations. Previous methods [14, 12] rely on single-step prediction strategies, which treat different levels of emotion variance independently and fail to capture the contextual dependencies between hierarchical emotion distributions. Additionally, current techniques often lack a structured mechanism to ensure that higher-level emotions influence lower-level prosodic variations, leading to inconsistencies in emotion expressiveness. Furthermore, integration with TTS remains another challenge, as most approaches are either model-specific or require reference audio, limiting their adaptability across different TTS architectures. Addressing these gaps, we propose a multi-step ED prediction framework that models hierarchical ED progressively, ensuring a more interpretable, flexible, and effective approach to speech emotion rendering and control.

3 Multi-step Prediction and Hierarchical Control of Emotion Intensity

We propose a novel approach, which supports the rendering of both single emotions and mixed emotions, that can be seamlessly integrated with various text-to-speech frameworks. Traditionally, speech databases label emotions at the utterance level, overlooking nuanced intensity variations within speech. To address this challenge, we automatically generate fine-grained, quantitative intensity labels, which serve as "soft labels" for speech generation models, eliminating the need for manual annotation. This method effectively enhances emotion control, enables mixed-emotion rendering, and can be readily adapted to speech generation frameworks, including text-to-speech and voice conversion.

3.1 Hierarchical Emotion Distribution (ED) Extractor

Built upon our previous studies [14, 12], the hierarchical emotion distribution (ED) extraction module integrates OpenSmile feature extractors [8] with pre-trained ranking functions at each segmental level to quantify emotion intensities in an utterance, as shown in Figure 2. Grounded in relative attributes [31], our method measures emotion prominence by treating emotion style as a speech attribute and ranking its presence relative to other emotions, enabling a structured and interpretable approach to hierarchical emotion quantification.

Specifically, we define the ranking function as:

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b \tag{1}$$

where $\mathbf{x}_i$, $\mathbf{w}$, and b denote the acoustic features of the i-th sample, weight vector, and bias, respectively. We optimize these parameters using a support vector machine objective for binary classification (e.g., Angry vs. Non-angry) [4] and normalize the outputs to the range $[0, 1]$, with larger values indicating stronger emotion intensity. This process enables continuous labeling of training data and the quantification of emotion intensity in unseen utterances during run-time.

Figure 2(b) illustrates our hierarchical ED extractor. We begin by segmenting the input audio into phoneme, word, and utterance levels using the Montreal Forced Aligner [28], and extracting an 88-dimensional feature set for each segment via OpenSMILE [8]. The pre-trained ranking functions then estimate an ED vector for each segment, where each element represents the intensity of a specific emotion. To ensure hierarchical consistency, we duplicate the utterance-level ED across all phonemes and replicate the word-level ED for the corresponding phonemes, as shown in Figure 2(a). These hierarchical ED vectors are subsequently incorporated into TTS training, which will be introduced in the next subsection.

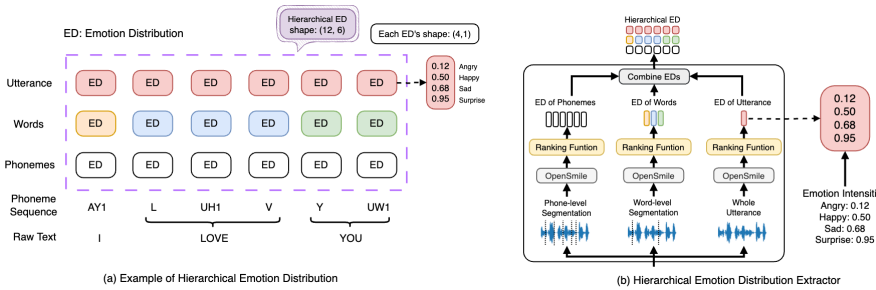

Figure 2: (a) Example of Hierarchical Emotion Distribution (ED) including EDs at levels of utterance, words, and phonemes; (b) Diagram of Hierarchical ED Extractor.

3.2 Multi-step Modeling of Hierarchical ED for TTS

We propose a multi-step strategy for modeling hierarchical emotion distribution (ED) to enable precise, multi-level control over emotion rendering in text-to-speech (TTS) synthesis. By progressively predicting ED at the utterance, word, and phoneme levels, our approach ensures that global emotional context guides local prosodic details. We explore two integration strategies to incorporate this multi-step hierarchical ED prediction into TTS frameworks.

3.2.1 External Integration

In the external integration approach ("External"), as shown in Figure 3(a), we enhance a model-agnostic TTS pipeline by integrating a hierarchical ED embedding after text processing. In this paper, we choose FastSpeech2 [33] as our TTS backbone. A text encoder converts phoneme sequences into linguistic embeddings, while a fully connected network transforms the hierarchical ED into an ED embedding. A variance adaptor then predicts pitch, duration, and energy, followed by a decoder that reconstructs the Mel-spectrogram using an L1 loss. This design effectively captures emotion intensity and improves prosody. Since the "External" framework does not inherently predict hierarchical ED, we incorporate a dedicated multi-step hierarchical ED prediction module. In this module, EDs are predicted sequentiallystarting from the utterance level, progressing to the word level, and finally to the phoneme levelwhile keeping the text encoder frozen as shown in Figure 3(b). This multi-step process allows a higher-level emotional context to guide finer, local prosodic adjustments.

3.2.2 Variation Adaptor Integration

The variation adaptor integration approach ("VA") integrates hierarchical ED modeling directly within the variance adaptor of FastSpeech2 [33]. This con-

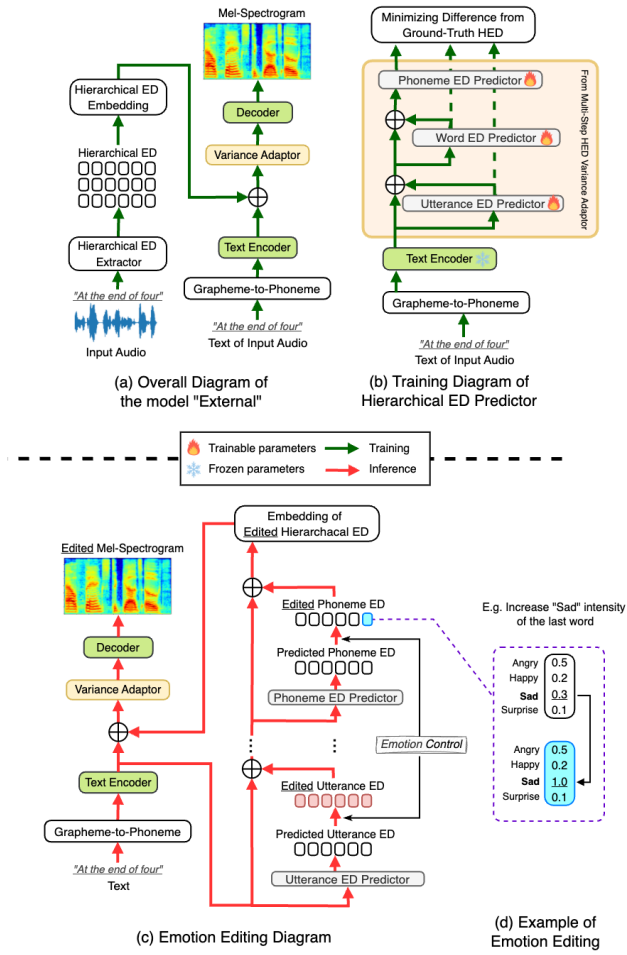

Figure 3: Training and Inference Diagrams of the proposed framework using external integration ("External"): (a) Overall diagram; (b) Training diagram of hierarchical emotion distribution (ED) predictor; (c) Emotion editing (inference) diagram (d) Example of emotion editing.

figuration extends the variance adaptor to jointly learn emotion representations and acoustic features, thereby tightly coupling emotion prediction with prosody generation. In contrast to the "External" setting, we train the linguistic encoder with a loss function that minimizes hierarchical ED differences. Specifically, we use a mean squared error (MSE) loss to reduce the discrepancy between the predicted and ground-truth EDs. Within the VA integration, ED prediction is also performed in a multi-step manner. The process begins with

predicting the utterance-level ED to establish the global emotional tone, which then informs the word-level prediction. Finally, these outputs are combined to generate phoneme-level ED, enabling fine-grained emotion control, as shown in Figure 4(b). This hierarchical, multi-step approach ensures that a broader emotional context effectively influences lower-level prosodic details.

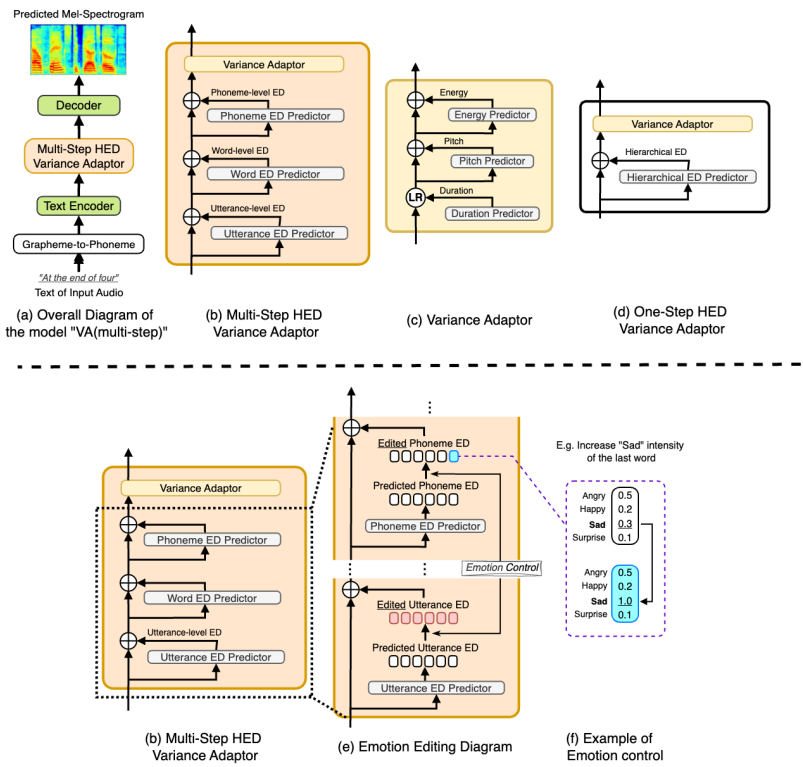

Figure 4: Training and Inference Diagrams of the proposed framework using variance adaptor integration ("VA"): (a) Overall diagram; (b) Diagram of sequential hierarchical emotion distribution (hierarchical ED) variance adaptor; (c) Diagram of variance adaptor; (d) Diagram of parallel hierarchical ED variance adaptor (e) Emotion editing (inference) diagram (f) Example of emotion editing.

3.3 Run-time Emotion Editing and Control

During run-time, our framework supports two primary tasks: (1) Emotion Control and (2) Emotion Editing. For emotion control, where only text input is available, our model predicts a hierarchical emotion distribution that aligns with the textual content. This allows users to control the emotion intensity

of individual speech segments, enabling fine-grained and quantifiable emotion rendering in synthesized speech. For the speech editing task, given an audio input and its corresponding transcript, our model extracts the hierarchical emotion distribution from the audio signal, enabling users control over emotion intensity for quantifiable emotion modification. In general, as depicted in Figure 3(c) and Figure 4(e), users are able to control the emotion rendering by adjusting the emotion distributions at three distinct levels, regardless of whether the ED is derived from audio or predicted from text.

4 Experimental Setup

We evaluated our system performance by conducting two experiments: (1) emotion prediction and (2) emotion editing. For emotion prediction, we train the TTS model on LibriTTS-R [18], a multi-speaker dataset containing approximately 580 hours of recordings from 2,306 speakers. Specifically, we utilized the "train-clean-100" and "train-clean-360" subsets for model training.

For the speech editing experiment, we used the Emotion Speech Dataset (ESD) [50, 51], which comprises over 29 hours of emotional speech in five categories—Neutral, Angry, Happy, Sad, and Surprise—from 20 speakers (10 native English and 10 Mandarin). We exclusively used the English recordings and the full training split for TTS model training. We also use the ESD dataset to train the hierarchical ED extractors ranking functions, where we randomly selected 100 samples per speaker and emotion, resulting in a total of 5,000 audio samples for hierarchical ED extractor training.

4.1 Model Architecture

We choose FastSpeech2 [33] as our backbone, which comprises a text encoder, variance adaptor, and decoder. We use a transformer [42]-based encoder to convert input phoneme sequences into linguistic embeddings. Variance adaptors predict hierarchical ED, duration, pitch, and energy. We utilize a transformer-based decoder to synthesize mel-spectrograms from these features. Our loss function combines the L1 loss on mel-spectrograms with the mean squared error for prosodic predictions in the variance adaptor. To address multi-speaker scenarios, we integrate speaker embeddings from Resemblyzer [43] into the encoder output. We adopt the Adam optimizer [16]. For TTS training, we use a batch size of 32 and perform 200,000 iterations over 48 hours on a single GPU. For text-based hierarchical ED prediction in "External", we conduct 100,000 iterations. The ED embedding layers consist of fully connected layers with a Tanh activation function. Finally, we employ HiFiGAN [19] as the vocoder, trained on the ESD and LibriTTS-R datasets.

4.2 Baselines Comparison

For emotion prediction experiments, we compared the model with our previous works [14, 12] ("single-step"), which predict EDs from text in a single-step manner. For example, in the "VA" setting, our proposed model progressively predicted utterance-, word-, and phoneme-level EDs (Figure 4(b)), whereas the baseline [14] predicted all segments concurrently (Figure 4(d)). For emotion editing experiments, we re-implemented MsEmoTTS [25] into the FastSpeech2 framework as the baseline ("MsEmoTTS") to ensure a fair comparison.

4.3 Evaluation Metrics

For objective evaluation, we calculated the Word Error Rate (WER) using Whisper[2] [32] to assess the system's robustness. To measure emotion similarity with the target, we evaluated spectral similarity using Mel-Cepstral Distortion (MCD) [21], prosody alignment through pitch and energy distortion, and duration deviation using Frame Disturbance (FD) [36].

For subjective evaluation, we conducted three listening experiments with 20 participants, each evaluating 210 synthesized samples. First, we conducted MUSHRA tests, where participants rated each speech sample on a scale from 0 to 100, with higher scores indicating better quality or greater similarity. The first test assessed speech naturalness, instructing participants to disregard noise and emotion. The second test evaluated emotion similarity, asking participants to rate the synthesized audio solely based on its emotional expressiveness while ignoring speech quality. Additionally, we conducted best-worst scaling (BWS) tests [17] to compare word-level emotion controllability between our model and the baseline. In BWStests, we adjusted the emotion intensity of three words per utterance to 0.0, 0.5, and 1.0, and evaluators selected the least and most expressive samples.

5 Experiments and Results

In this section, we present the experimental results for two tasks: (1) Emotion Prediction and (2) Emotion Editing. For Emotion Prediction, we evaluated our models based on speech quality and emotional expressiveness, supplemented by qualitative analysis. For Emotion Editing, we assessed the controllability of emotions, measuring how effectively the model modifies and adjusts emotion intensity for different speech segmental levels.

[2]Whisper Large: https://github.com/openai/whisper.

5.1 Experiments with Emotion Prediction

We compared synthesized audio samples across seven different conditions, as detailed in Table 1 and Table 2. These tables organize these conditions into three key categories: "GT or Pred", "TTS Model", and "Pred Mode":

- **GT or Pred**: This column specifies whether the hierarchical emotion distributions (EDs) are obtained from ground-truth data (GT) or predicted from text.
- **TTS Model**: This column indicates the TTS model used. VA refers to the model utilizing the Single-Step hierarchical ED Variance Adaptor (Figure 4(d)), while VA (Multi-Step) corresponds to the model employing the Multi-Step hierarchical ED Variance Adaptor (Figure 4(b)).
- **Pred Mode**: This column describes how the hierarchical ED is predicted. It is either generated progressively from longer to shorter segments ("Multi-Step") or in parallel for all segments at once ("Single-Step").

This configuration allows for a structured comparison of how different hierarchical ED configurations and prediction strategies impact synthesis quality and emotion expressiveness.

Table 1: Speech Quality Test Results: MUSHRA naturalness scores with 95% confidence interval and Word Error Rate (WER). The column "GT or Pred" indicates whether we use ground-truth hierarchical ED ("GT") or a text-predicted version. In the "TTS Model" column, "VA" and "VA(Multi-Step)" denote the TTS models employing the Single-Step hierarchical ED Variance Adaptor (Figure 4(d)) and the Multi-Step hierarchical ED Variance Adaptor (Figure 4(b)), respectively. Finally, the "Pred Mode" column specifies whether we predict the hierarchical ED sequentially from longer to shorter segments (Multi-Step) or in parallel for all segments (Single-Step).

Hierarchical ED			Speech Quality	
GT or Pred	TTS Model	Pred Mode	MUSHRA (↑)	WER (↓)
— Ground-Truth Speech Samples —			$79.4_{\pm 1.9}$	2.16
GT	External	-	$61.6_{\pm 2.2}$	3.37
GT	VA	-	$57.5_{\pm 2.6}$	3.11
GT	VA(Multi-Step)	-	$\mathbf{62.2}_{\pm 2.3}$	**2.48**
Predicted	External	Single-Step	$50.7_{\pm 2.4}$	3.80
Predicted	External	Multi-Step	$\mathbf{54.0}_{\pm 2.3}$	**3.25**
Predicted	VA	Single-Step	$52.2_{\pm 2.6}$	4.61
Predicted	VA(Multi-Step)	Multi-Step	$\mathbf{53.2}_{\pm 2.4}$	**2.45**

Table 2: Emotion Expressiveness Test Results with 95% confidence interval: MUSHRA similarity scores, Mel-Cepstral Distortion (MCD), Pitch/Energy Distortion (Pitch/Energy), and Frame Disturbance (FD). The column "GT or Pred" indicates whether we use ground-truth hierarchical ED ("GT") or a text-predicted version. In the "TTS Model" column, "VA" and "VA(Multi-Step)" denote the TTS models employing the Single-Step hierarchical ED Variance Adaptor (Figure 4(d)) and the Multi-Step hierarchical ED Variance Adaptor (Figure 4(b)), respectively. Finally, the "Pred Mode" column specifies whether we predict the hierarchical ED progressively from longer to shorter segments (Multi-Step) or in parallel for all segments (Single-Step).

Hierarchical ED			Emotion Expressiveness				
GT or Pred	TTS Model	Pred Mode	MUSHRA (↑)	MCD (↓)	Pitch (↓)	Energy (↓)	FD (↓)
GT	External	-	**61.9**$_{\pm 2.1}$	5.88$_{\pm 0.10}$	15.6$_{\pm 1.0}$	0.363$_{\pm 0.022}$	24.8$_{\pm 3.4}$
GT	VA	-	55.8$_{\pm 2.6}$	6.48$_{\pm 0.23}$	16.1$_{\pm 1.1}$	0.386$_{\pm 0.023}$	25.5$_{\pm 3.6}$
GT	VA(Multi-Step)	-	61.9$_{\pm 2.1}$	**5.62**$_{\pm 0.11}$	**15.5**$_{\pm 1.1}$	**0.348**$_{\pm 0.020}$	**22.4**$_{\pm 2.9}$
Predicted	External	Single-Step	47.2$_{\pm 2.3}$	7.59$_{\pm 0.14}$	18.2$_{\pm 1.1}$	0.438$_{\pm 0.027}$	46.3$_{\pm 6.5}$
Predicted	External	Multi-Step	**51.9**$_{\pm 2.2}$	**6.89**$_{\pm 0.12}$	**16.9**$_{\pm 1.2}$	**0.409**$_{\pm 0.024}$	**42.6**$_{\pm 4.7}$
Predicted	VA	Single-Step	48.2$_{\pm 2.5}$	7.23$_{\pm 0.20}$	**16.7**$_{\pm 1.0}$	0.426$_{\pm 0.025}$	**41.4**$_{\pm 4.7}$
Predicted	VA(Multi-Step)	Multi-Step	**49.1**$_{\pm 2.2}$	**6.91**$_{\pm 0.12}$	17.2$_{\pm 1.1}$	**0.416**$_{\pm 0.025}$	46.3$_{\pm 5.1}$

5.1.1 Speech Quality Evaluation

Table 1 summarizes the results of the MUSHRA and WER tests. With ground-truth hierarchical ED, the variance adapter with multi-step emotion modeling (VA (Multi-Step)) consistently outperforms the single-step approach (VA), achieving higher MUSHRA scores and lower WER. We also observe that, when using predicted hierarchical ED, the multi-step models significantly improve both speech naturalness and intelligibility compared to the single-step models. These findings suggest that aligning the EDs of shorter segments with those of longer segments is crucial for enhancing overall speech quality.

5.1.2 Emotion Expressiveness Evaluation

We further assess emotion expressiveness by conducting MUSHRA tests on emotional similarity and computing multiple objective prosody-related metrics, including Mel-Cepstral Distortion (MCD), Pitch/Energy Distortion, and Frame Disturbance (FD). As shown in Table 2, our proposed multi-step hierarchical ED prediction consistently achieves higher MUSHRA emotional similarity scores and lower distortion values across all objective metrics for both ground-truth and predicted hierarchical EDs. These results highlight the effectiveness of the multi-step scheme in enhancing emotion expressiveness and improving alignment with ground-truth emotions. However, we observe that in the "VA" setting, the multi-step scheme does not outperform the single-step approach in Pitch and FD. This discrepancy may stem from error accumulation across different levels of ED prediction (utterance, word, and phoneme levels). Compared to the "External" setting, where the linguistic encoder

remains independent to emotion prediction, the "VA" setting incorporates hierarchical emotion distribution difference loss in training. This joint training may increase the models sensitivity to ED variations in longer segments, potentially leading to greater fluctuations in prosody-related metrics.

5.1.3 Analysis on Hierarchical ED Prediction

We further analyzed the predicted hierarchical emotion distribution (ED). Table 3 presents the mean absolute difference between the predicted and ground-truth values for each segment. The column "Longer Segments" denotes the longer segments used to predict shorter segments; "GT" indicates that ground-truth values were employed. For example, under the "GT" condition, we used the ground-truth utterance-level ED to predict the word-level ED, whereas under the "Predicted" condition, we utilized the predicted utterance-level ED.

Table 3: Mean Absolute Difference of Hierarchical ED: differences between the predicted and the ground-truth hierarchical ED values. The column Longer "Longer Segments" denotes the longer segments used to predict shorter segments; "GT" indicates that ground-truth values were employed. For example, under the "GT" condition, we used the ground-truth utterance-level ED to predict the word-level ED, whereas under the "Predicted" condition, we utilized the predicted utterance-level ED.

Hierarchical ED Condition			Hierarchical ED Difference			
TTS Model	Pred Mode	Longer Segments	Phonemes	Words	Utterance	Avg.
External	Single-Step	-	0.1333	0.1283	0.0594	0.1070
External	Multi-Step	Predicted	0.1345	0.1297	**0.0587**	0.1077
External	Multi-Step	GT	**0.1214**	**0.1281**	0.0587	**0.1028**
VA	Single-Step	-	0.1358	0.1294	**0.0599**	0.1084
VA(Multi-Step)	Multi-Step	Predicted	0.1356	0.1298	0.0601	0.1085
VA(Multi-Step)	Multi-Step	GT	**0.1230**	**0.1272**	0.0601	**0.1034**

Table 3 summarizes our results. We did not observe significant differences between single-step and multi-step predictions, despite substantial differences in synthesized audio performances (see Tables 1 and 2). These results suggest that our prediction modules not only reduce hierarchical ED differences but also generate emotion representations consistent with the textual emotional content in the audio domain. More importantly, when comparing "Predicted" and "GT" in the multi-step mode, we found error accumulation in ED prediction, evidenced by a smaller gap at the word level and an increased gap at the phoneme level. These findings suggest that our model prioritizes the dependency of EDs across segments over mere ED differences, which explains the improved speech naturalness (Table 1) and only marginally better emotion expressiveness (Table 2).

Next, we analyzed the predicted word- and phoneme-level EDs derived from different utterance-level EDs. Specifically, we systematically varied the

intensity of a single emotion, setting it to 1.0 while keeping the intensities of all other emotions at 0.0. We then visualized the resulting ED distributions as histograms for each segment in Figure 5, where each row represents the intensified emotion and each column corresponds to a segment. We utilized a TTS model trained on the ESD dataset to highlight the impact of emotional variations in speech synthesis. From Figure 5, we observe that both word- and phoneme-level EDs exhibit a positive correlation with their respective utterance-level ED values, with the correlation being stronger at the word level. This suggests that emotion propagation is more consistent across words than phonemes. Additionally, we find that word-level anger and surprise intensities display a notably stronger inter-correlation compared to other emotions, indicating that these two emotions may share similar prosodic and acoustic patterns at the word level. This observation aligns with psychological studies suggesting that anger and surprise often exhibit overlapping acoustic characteristics, such as increased pitch and energy.

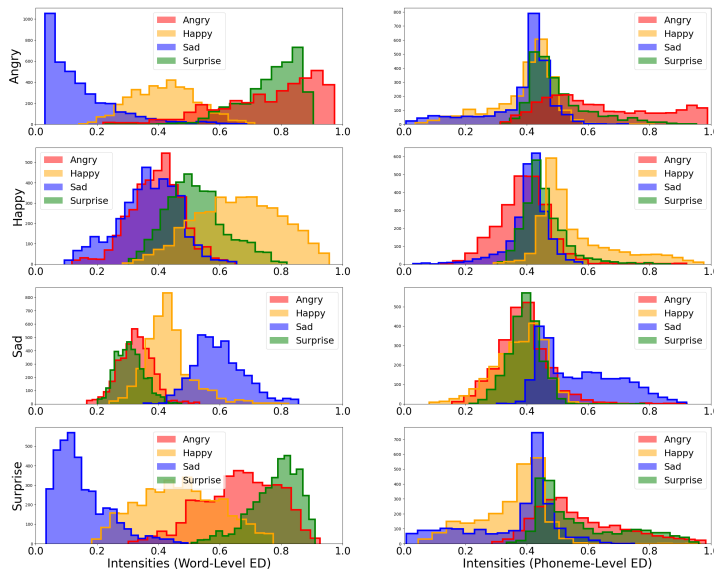

Figure 5: Histograms of word-level and phoneme-level emotion distributions (EDs) for various intensified utterance-level EDs. Each row corresponds to an intensified emotion, and each column corresponds to a segment.

5.2 Experiments with Emotion Editing

Following [12], we evaluated our models' emotion controllability on the ESD dataset through subjective evaluations and objective analysis. We first con-

ducted best-worst scaling (BWS) tests [17] to compare word-level emotion controllability between our model and the baseline MsEmoTTS [25]. As presented in Table 4, our model exhibited a stronger tendency than MsEmoTTS to select the least expressive sample at low intensity and the most expressive sample at high intensity across all five emotions. This trend was especially pronounced for Sad and Surprise emotions, where the distinction between intensity levels was more evident. These results demonstrate our model's ability to capture fine-grained variations in emotional intensity, ensuring more precise and consistent emotion rendering control compared to the baseline MsEmoTTS.

Table 4: Best-Worst Scaling (BWS) Test Result: The value represents evaluator preferences (%), with red and blue indicating the heatmap for the least expressive and most expressive audio, respectively.

		Hierarchical ED				MsEmoTTS			
		Ang	Hap	Sad	Sur	Ang	Hap	Sad	Sur
Least	0.0	79	63	67	81	42	32	21	33
	0.5	0	28	14	14	47	47	54	42
	1.0	21	9	19	5	11	21	25	25
Most	0.0	11	18	16	7	11	12	30	18
	0.5	16	7	9	7	26	16	23	30
	1.0	74	75	75	86	63	72	47	53

We further validated controllability across utterance, word, phoneme, and word-and-phoneme levels. We incremented emotion intensity from 0.0 to 1.0 and computed prosodic features such as duration and the mean/standard deviation of pitch and energy (Figure 6). Because duration values vary between levels, we standardized them prior to visualization. Prior literature [35] correlates these features with emotion intensity; for example, sadness is associated with a slower speaking rate and lower pitch and energy values. We analyzed the ESD dataset [50] to examine these acoustic-emotion relationships. A red background indicates a negative trend, while blue is a positive trend with increasing intensity. Our model followed these expected trends, showing a positive correlation between happiness and mean pitch and a negative correlation between sadness and pitch. Additionally, editing both word and phoneme-level emotions produced significant prosodic changes, with the standard deviation of pitch at the utterance level aligning with our expectations.

Figure 7 shows the spectrograms of synthesized audio samples with varying emotion intensities. We display pitch and energy contours (blue and green lines, respectively), noting that the y-axis for the energy contours is not relevant. Figures 7(a) and (b) present utterance and wordlevel intensity control. Each row corresponds to a different emotion, with the first column depicting acoustic features at an emotion intensity of 0.0, and the second column at 1.0.

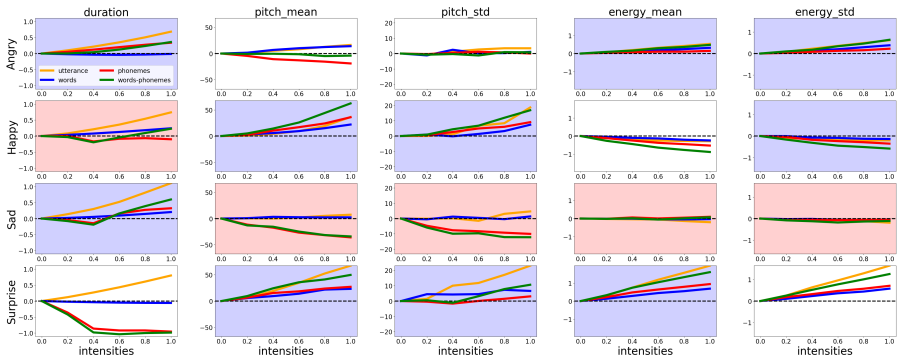

Figure 6: The illustration of prosodic variants with intensity changes. The red background represents the expected negative trend, the blue indicates the expected positive trend, both summarized from the ESD dataset.

In (b), the three highlighted areas indicate regions where the intensity has been modified. For both segments, for anger, we observe more pronounced energy spikes, at higher intensities. In happiness, pitch and energy patterns are similar, with higher pitch values at an intensity of 1.0. Sadness prolongs duration and stabilizes pitch contour as intensity increases. For surprise, we note a rise in both pitch contours and energy spikes. These results align with Figure 6 and demonstrate that our model's ability to manipulate duration and pitch/energy according to emotion intensity.

6 Conclusion

We present a multi-step prediction framework for hierarchical emotion distribution (ED), enabling multi-level control of emotion rendering in speech synthesis. By modeling ED at the utterance, word, and phoneme levels progressively, our approach ensures that higher-level emotional context influences lower-level prosody, resulting in more natural and expressive speech. We integrate ED into TTS systems through two strategies: embedding it within the variance adaptor of FastSpeech2 and incorporating it as an external module for other non-autoregressive TTS models, making our method flexible and widely applicable. At runtime, users can quantitatively control emotion intensity, enhancing the interpretability and adaptability of emotional speech synthesis. Objective and subjective evaluations demonstrate that our approach significantly improves speech quality, expressiveness, and controllability. In future work, we will extend our proposed method to additional languages, varied voice qualities, and more diverse emotional datasets, further exploring cross-lingual robustness and broadening the applicability of multi-level emotion intensity prediction.

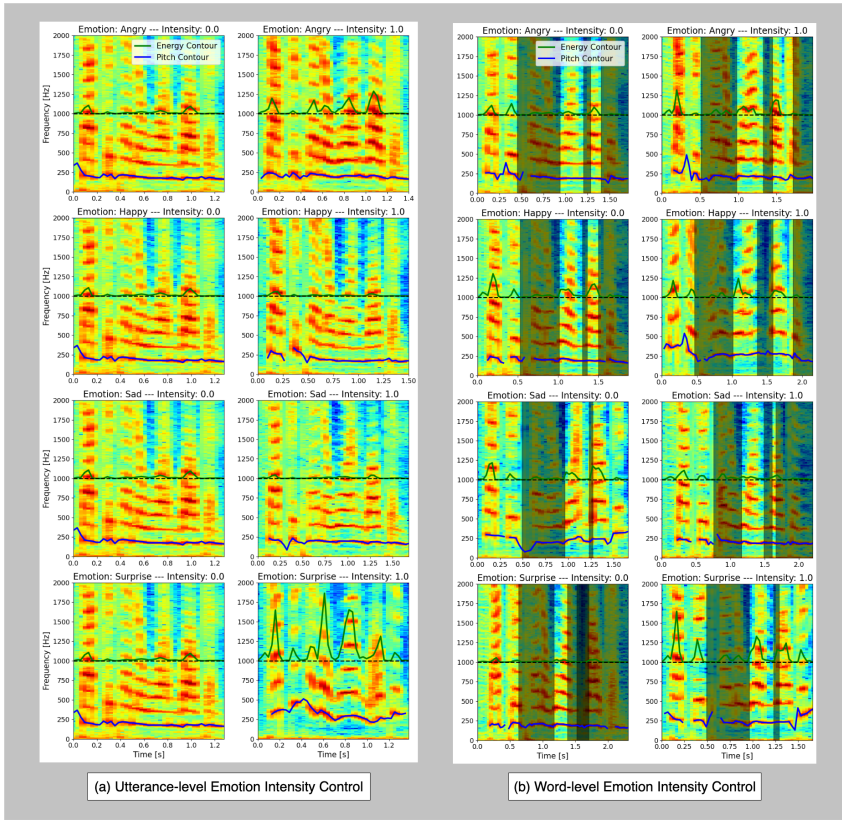

Figure 7: Spectrograms of synthesized audio samples across different emotion intensities with pitch (blue) and energy (green) contours: the y-axis for energy contours are not relevant. (a) Utterance-level Emotion Intensity Control (b) Word-level Emotion Intensity Control.

References

[1] C. Busso, S. Lee, and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection", *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4), 2009, 582–96, DOI: 10.1109/TASL.2008.2009578.

[2] 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 *Interspeech 2024*, 2024, 1810–4, DOI: 10.21437/Interspeech.2024-398.

[3] 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", 2024, https://arxiv.org/abs/2411.02625.

[4] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, 20(3), 1995, 273–97.

[5] Y. Cui, X. Wang, Z. Zhao, W. Zhou, and H. Chen, "CASEIN: Cascading Explicit and Implicit Control for Fine-grained Emotion Intensity Regulation", 2023, arXiv: 2307.00020 [cs.SD].

[6] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern recognition*, 44(3), 2011, 572–87.

[7] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, 44(3), 2011, 572–87, ISSN: 0031-3203, DOI: https://doi.org/10.1016/j.patcog.2010.09.020, https://www.sciencedirect.com/science/article/pii/S0031320310004619.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor", in, January 2010, 1459–62, DOI: 10.1145/1873951.1874246.

[9] J. Hirschberg, "Pragmatics and intonation", *The handbook of pragmatics*, 2006, 515–37.

[10] J. Hirschberg and G. Ward, "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English", *Journal of Phonetics*, 20(2), 1992, 241–51, ISSN: 0095-4470, DOI: https://doi.org/10.1016/S0095-4470(19)30625-4, https://www.sciencedirect.com/science/article/pii/S0095447019306254.

[11] 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 *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, 6317–21, DOI: 10.1109/ICASSP43922.2022.9747098.

[12] S. Inoue, K. Zhou, S. Wang, and H. Li, "Fine-Grained Quantitative Emotion Editing for Speech Generation", *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, 1–6, https://api.semanticscholar.org/CorpusID:268248771.

[13] S. Inoue, K. Zhou, S. Wang, and H. Li, "Hierarchical Control of Emotion Rendering in Speech Synthesis", 2025, arXiv: 2412.12498 [cs.SD], https://arxiv.org/abs/2412.12498.

[14] S. Inoue, K. Zhou, S. Wang, and H. Li, "Hierarchical Emotion Prediction and Control in Text-to-Speech Synthesis", in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, 10601–5, DOI: 10.1109/ICASSP48485.2024.10445996.

[15] X. Jing, K. Zhou, A. Triantafyllopoulos, and B. W. Schuller, "Enhancing Emotional Text-to-Speech Controllability with Natural Language Guidance through Contrastive Learning and Diffusion Models", *arXiv preprint arXiv:2409.06451*, 2024.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", 2017, arXiv: 1412.6980 [cs.LG].

[17] S. Kiritchenko and S. Mohammad, "Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ed. R. Barzilay and M.-Y. Kan, Vancouver, Canada: Association for Computational Linguistics, July 2017, 465–70, DOI: 10.18653/v1/P17-2074, https://aclanthology.org/P17-2074.

[18] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus", *arXiv preprint arXiv:2305.18802*, 2023.

[19] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", 2020, arXiv: 2010.05646 [cs.SD].

[20] S. R. Krothapalli and S. G. Koolagudi, "Emotion Recognition Using Prosodic Information", in, *Emotion Recognition using Speech Features*, New York, NY: Springer New York, 2013, 79–91, ISBN: 978-1-4614-5143-3, DOI: 10.1007/978-1-4614-5143-3_5, https://doi.org/10.1007/978-1-4614-5143-3_5.

[21] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment", in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Vol. 1, 1993, 125–128 vol.1, DOI: 10.1109/PACRIM.1993.407206.

[22] Z. KUN, "Emotion modelling for speech generation", 2022.

[23] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis", in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[24] S. Lei, Y. Zhou, L. Chen, J. Hu, Z. Wu, S. Kang, and H. Meng, "Towards Multi-Scale Speaking Style Modelling with Hierarchical Context Information for Mandarin Speech Synthesis", 2022, arXiv: 2204.02743 [cs.SD].

[25] Y. Lei, S. Yang, X. Wang, and L. Xie, "MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis", 2022, arXiv: 2201.06460 [cs.SD].

[26] L. Leinonen, T. Hiltunen, I. Linnankoski, and M.-L. Laakso, "Expression of emotional–motivational connotations with a one-word utterance", *Journal of the Acoustical Society of America*, 102(3), 1997, 1853–63, https://doi.org/10.1121/1.420109.

[27] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis", 2022, arXiv: 2109.06733 [cs.SD].

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi", in *Proc. Interspeech 2017*, 2017, 498–502, DOI: 10.21437/Interspeech.2017-1386.

[29] H. Ming, D.-Y. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion.", in *Interspeech*, 2016, 2453–7.

[30] Y. Oh, J. Lee, Y. Han, and K. Lee, "Semi-supervised learning for continuous emotional intensity controllable speech synthesis with disentangled representations", 2023, arXiv: 2211.06160 [eess.AS].

[31] D. Parikh and K. Grauman, "Relative attributes", in *2011 International Conference on Computer Vision*, IEEE, 2011, 503–10.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision", 2022, arXiv: 2212.04356 [eess.AS], https://arxiv.org/abs/2212.04356.

[33] 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", 2022, arXiv: 2006.04558 [eess.AS].

[34] E. Rodero, "Intonation and Emotion: Influence of Pitch Levels andăContour Type on Creating Emotions", *Journal of Voice*, 25(1), 2011, e25–e34, ISSN: 0892-1997, DOI: https://doi.org/10.1016/j.jvoice.2010.02.002, https://www.sciencedirect.com/science/article/pii/S0892199710000378.

[35] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends", *Communications of the ACM*, 61(5), 2018, 90–9.

[36] B. Sisman, G. Lee, H. Li, and K. C. Tan, "On the analysis and evaluation of prosody conversion techniques", in *2017 International Conference on Asian Language Processing (IALP)*, 2017, 44–7, DOI: 10.1109/IALP.2017.8300542.

[37] J. Snedeker and J. Trueswell, "Using prosody to avoid ambiguity: Effects of speaker awareness and referential context", *Journal of Memory and Language*, 48 (January), January 2003, 103–30, DOI: 10.1016/S0749-596X(02)00519-3.

[38] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis", *arXiv preprint arXiv:2106.15561*, 2021.

[39] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis", 2023, arXiv: 2306.00648 [cs.SD].

[40] A. Triantafyllopoulos and B. W. Schuller, "Expressivity and Speech Synthesis", *arXiv preprint arXiv:2404.19363*, 2024.

[41] A. Triantafyllopoulos, B. W. Schuller, G. ymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André, *et al.*, "An overview of affective speech synthesis and conversion in the deep learning era", *Proceedings of the IEEE*, 2023.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", 2017, arXiv: 1706.03762 [cs.CL].

[43] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification", 2020, arXiv: 1710.10467 [eess.AS].

[44] S. Wang, J. Guonason, and D. Borth, "Fine-Grained Emotional Control of Text-to-Speech: Learning to Rank Inter- and Intra-Class Emotion Intensities", in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5, DOI: 10.1109/ICASSP49357.2023.10097118.

[45] A. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas", *Behavior research methods*, 45 (February), February 2013, DOI: 10.3758/s13428-012-0314-x.

[46] Y. Xu, "Speech prosody: A methodological review", *Journal of Speech Sciences*, 1(1), 2011, 85–115.

[47] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iEmoTTS: Toward Robust Cross-Speaker Emotion Transfer and Control for Speech Synthesis based on Disentanglement between Prosody and Timbre", 2023, arXiv: 2206.14866 [eess.AS].

[48] K. Zhou, B. Sisman, C. Busso, B. Ma, and H. Li, "Mixed-EVC: Mixed Emotion Synthesis and Control in Voice Conversion", 2023, arXiv: 2210.13756 [eess.AS].

[49] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data", in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, 230–7.

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

[51] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset", in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 920–4.

[52] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech Synthesis with Mixed Emotions", 2022, arXiv: 2208.05890 [cs.CL].

[53] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion Intensity and its Control for Emotional Voice Conversion", *IEEE Transactions on Affective Computing*, 14(1), 2023, 31–48, DOI: 10.1109/TAFFC.2022.3175578.

[54] 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.

[55] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling Emotion Strength with Relative Attribute for End-to-End Speech Synthesis", in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, 192–9, DOI: 10.1109/ASRU46091.2019.9003829.