university of groningen

campus fryslân

Towards Fine-Grained Emotional Modulation in FastSpeech 2 with Hierarchical Emotion Distributions

Qiyan Huang

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Towards Fine-Grained Emotional Modulation in FastSpeech 2 with Hierarchical Emotion Distributions

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Asst. Prof. Dr. Vass Verkhodanova (Voice Technology, University of Groningen)
with the second reader being
Asst. Prof. Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Qiyan Huang (S-5858895)

June 11, 2025

Acknowledgements

I wish to express my sincere appreciation to all those who contributed to the completion of this thesis.

First and foremost, I am deeply grateful to my supervisor, Vass Verkhodanova,for her kindly guidance and support throughout this research journey. The weekly discussions with Vass proved invaluable in shaping my analytical approach and research methodology. Her detailed feedback and critical insights were instrumental in clarifying my research objectives and strengthening the theoretical foundation of this work. Her reviews and constructive suggestions during the revision process significantly enhanced the coherence, organization, and scholarly quality of this manuscript.

I extend my heartfelt thanks to all participants who voluntarily contributed to the questionnaire survey. Their thoughtful responses provided essential subjective perspectives that enriched the technical analysis and added depth to the overall research findings.

I would also like to acknowledge the dedication and perseverance I maintained throughout the demanding phases of experimentation and writing. The extensive hours devoted to model optimization, data analysis, and manuscript refinement represent not only my academic commitment but also a significant milestone in personal and intellectual development.

Finally, I am grateful to my classmates and friends for their constant support and assistance throughout this journey.

Abstract

Emotional speech synthesis has made substantial progress; however, interpretable and fine-grained prosody controlremains a persistent challenge. Existing systems often rely on global emotion labels or latent style embeddings, which limits precise temporal manipulation of emotional expression.

This thesis introduces a novel approach to emotional prosody control by integrating phoneme-aligned Hierarchical Emotion Distributions (HED) into the non-autoregressive FastSpeech 2 architecture. The method enables interpretable emotion conditioning through injecting 12-dimensional HED vectors after the variance adaptor, supported by a gradual training strategy for stable convergence.

Experiments, conducted using the English subset of the Emotional Speech Dataset (ESD), employed multiple evaluation settings. These included sentence- and phoneme-level acoustic analysis, inference-time intensity manipulation, and perceptual testing via Best-Worst Scaling (BWS). Models were compared across emotion categories and training stages to assess control effectiveness and robustness.

Results demonstrate that HED conditioning yields consistent, emotion-specific prosodic patterns with clearly distinguishable pitch and energy trajectories. Furthermore, inference-time manipulation of HED vectors results in predictable changes in emotional intensity, confirming the proposed system's controllability. Subjective ratings align with acoustic findings, showing listener preference for HED-guided outputs.

This research contributes a structured and interpretable framework for emotional speech synthesis, advancing the controllability of non-autoregressive TTS. This work supports future applications in expressive voice technologies, virtual agents, and human-computer interaction.

Contents

1 Introduction

Speech synthesis has experienced a dramatic evolution over the past few decades. Traditional techniques such as formant-based synthesis and unit selection systems offered early attempts at generating intelligible speech, but these approaches suffered from limited flexibility and poor generalization to unseen contexts. The advent of deep learning brought a transformative shift to the field. WaveNet(van den Oord et al., 2016) demonstrated that autoregressive neural networks could produce raw waveforms with unprecedented naturalness. Subsequently, Tacotron and Tacotron 2 (Wang et al., 2017)introduced end-to-end frameworks that directly mapped text to spectrograms, significantly streamlining the synthesis pipeline. More recently, VITS (Kim, Kong, & Son, 2021)integrated variational inference, adversarial training, and normalizing flows to simultaneously achieve high-quality synthesis and efficient inference.

While these neural TTS models have greatly improved the naturalness and fluency of synthetic speech, they still fall short in providing explicit, interpretable, and fine-grained control over prosodic and emotional expression. Current systems, including Tacotron 2 and VITS, often rely on global emotion embeddings or sentence-level emotion labels, which apply a uniform emotional style to an entire utterance. This uniformity limits the ability to express localized emotional variations, especially at the phoneme or word level. Techniques such as style tokens and variance encoders have attempted to capture prosodic variation, but the resulting latent representations tend to be entangled, opaque, and difficult to manipulate in a controlled or quantifiable manner(Du & Yu, 2021)(Tits, 2022). Meanwhile, user-driven speech editing frameworks (Morrison et al., 2021) (Tae, Kim, & Kim, 2022) offer partial solutions but frequently lack consistency in controlling emotion intensity or prosodic dynamics across utterances.

These limitations stand in contrast to findings from both perceptual and linguistic studies.Scherer (2003) demonstrated that emotional cues can be reliably perceived from segments as short as a single syllable, indicating the importance of local emotional information.Xu (2019) further proposed that global prosodic contours, such as intonation and rhythm, emerge from the sequential implementation of locally defined pitch targets at the syllable or phoneme level. These insights suggest that emotional expression should not be confined to the utterance level, but rather should be realized through fine-grained prosodic variation. In conversational speech, the ability to modulate emotional expression at the phoneme level is critical for conveying contrastive focus, emphasis, or irony. For example, subtle shifts in pitch or timing on the word “really” can convey surprise, doubt, or sarcasm depending on the prosodic configuration.

In response to these insights, recent efforts have explored hierarchical prosody modeling frameworks(Hsu et al., 2018)(Skerry-Ryan et al., 2018), which aim to model prosodic variation across multiple linguistic levels. However, the latent prosodic features learned by these models often remain uninterpretable and lack alignment with specific emotional dimensions. As a result, these systems are still unable to provide transparent and consistent emotional control, particularly at fine temporal resolutions.

Building on recent advances in hierarchical prosody modeling and structured emotional editing(Inoue, Zhou, Wang, & Li, 2024a), this thesis investigates the integration of phoneme-level Hierarchical

Emotion Distributions (HED) into the FastSpeech 2 architecture. Unlike previous work, which typically applies emotion conditioning at the utterance or word level, the proposed approach injects HED vectors at the phoneme level, enabling fine-grained emotional modulation. The injection mechanism and training strategy are designed to explore whether such localized conditioning can produce interpretable and controllable emotional prosody in non-autoregressive TTS.

1.1 Research Questions and Hypotheses

This study examines whether phoneme-aligned HED vectors, which encode emotion intensity over time based on low-level acoustic features, can serve as an effective and interpretable control mechanism for emotional prosody generation in FastSpeech 2.

To address this objective, three sub-questions and their corresponding hypotheses are formulated as follows:

Sub-question 1: Emotion specific modulation does conditioning FastSpeech 2 with phoneme-level HED vectors corresponding to distinct emotional categories result in systematically distinguishable prosodic patterns at the utterance level?

Hypothesis 1: HED vectors corresponding to different emotional categories will produce utterances with statistically and perceptually distinct prosodic profiles, reflected in pitch(Yoon, 2024) and energy contours aligned with established emotion-expression patterns.

Sub-question 2: Training stage robustness Does the effect of HED-based emotional conditioning become more stable and consistent as training progresses?

Hypothesis 2: It is hypothesized that as the model continues to train, its ability to map HED vectors to emotional prosody will improve. This will be reflected by more stable acoustic trends across training checkpoints, reduced within-class variability, and increasingly smooth F0 and energy contours for each emotion.

Sub-question 3: Emotional intensity controllability Under a fixed sentence and emotion category, do different types of phoneme-level HED vectors including zero vectors, full-one vectors, and reference vectors lead to prosodically distinct outputs reflecting varying levels of emotional intensity?

Hypothesis 3: It is hypothesized that, under fixed text and emotion category, different types of HED vectors will result in clearly different levels of emotional intensity in synthesized prosody.Specifically, zero vectors are expected to yield neutral prosody, full-one vectors to produce overly expressive contours, and reference vectors to reflect balanced emotional intensity.

1.2 Terminology Definition

This section defines the three core concepts that guide the evaluation of phoneme-level emotional conditioning in this study:

- Interpretable: Refers to a clear and consistent mapping between the injected phoneme-level HED vectors and the resulting prosodic features (e.g., F0, energy). In this study, interpretability is assessed by examining whether systematic variations in HED inputs lead to predictable and differentiable changes in prosodic patterns across emotions and intensity levels.
- Consistent: Refers to the stability of emotion-specific prosodic patterns over the course of model training. A consistent system exhibits reduced variability and converges to recognizable acoustic trends for each emotional category.
- Emotional Intensity Controllable: Refers to the model's ability to modulate the strength of emotional expression by systematically manipulating the emotional input vector.(Inoue, Zhou, Wang, & Li, 2024b)Similar to the approach used in diffusion-based models like EmoDiff(Guo, Du, Chen, & Yu, 2023), where emotion strength is continuously varied via soft-label interpolation, the HED-based method aims to enable nuanced and scalable control over prosodic intensity.

1.3 Research Contribution

Despite increasing interest in controllable emotional TTS, few studies have explored phoneme-level emotional conditioning in non-autoregressive architectures. This research addresses this gap by investigating how phoneme-aligned Hierarchical Emotion Distributions (HED) can guide fine-grained prosody generation.

By integrating HED into the FastSpeech 2 architecture, this study introduces a structured and interpretable approach for emotional prosody control. It further examines training dynamics and demonstrates the feasibility of scaling emotional intensity via HED vector manipulation.

Although emotion vectors are injected at the phoneme level, prosodic expressivity is evaluated at the sentence level to reflect holistic listener perception. The results provide insights into how localized emotion cues influence global expressivity, contributing to more controllable, interpretable, and inclusive emotional speech synthesis systems.

2 Literature Review

2.1 Global and Latent Control Methods

Recent advances in expressive text-to-speech (TTS) synthesis have significantly improved speech naturalness and prosodic variation, largely due to the integration of latent prosody and global emotion modeling techniques. A common paradigm involves learning a fixed-size latent vector that encodes global prosodic characteristics such as pitch, duration, and energy. These vectors are typically derived from either reference speech or latent generators and are conditioned on the decoder to guide prosodic realization.

For instance, DiffProsody (Oh, Lee, & Lee, 2024) introduces a diffusion-based latent prosody generator that synthesizes prosodic vectors without relying on reference audio. By leveraging a denoising diffusion GAN (DDGAN), the model captures complex prosodic patterns from text and speaker embeddings while achieving significant inference acceleration compared to autoregressive predictors. However, the latent vectors produced by this model lack semantic interpretability and editability, which restricts their applicability in user-controllable emotional modulation.

Similarly, Hierarchical Prosody Modeling (Jiang et al., 2024) enhances zero-shot expressiveness by incorporating a hierarchical adaptor that operates across multiple resolutions,such as frame, phoneme, and word. Combined with a diffusion-based pitch predictor and a global speaker embedding, the model achieves fine-grained prosody manipulation and speaker generalization. Nevertheless, the prosodic features are fused implicitly with the linguistic content, offering little transparency or direct control over affective attributes.

To address the limitations of global emotion vectors, EmoKnob(Chen, Chen, & Hirschberg, 2024)proposes a few-shot emotion control method that manipulates speaker embeddings to adjust emotional intensity. By computing the difference between neutral and emotional embeddings, the model derives an emotion direction vector that can be scaled and applied to new utterances. While EmoKnob supports open-ended emotion prompts and fine-grained control over expressivity, the underlying representation remains latent and entangled, lacking interpretability and phoneme-level precision.

In summary, although latent vector based approaches have advanced the generation of expressive synthetic speech, their control mechanisms are largely restricted to the utterance or speaker embedding level. These methods seldom provide explicit control over prosodic correlates such as pitch or energy, and the opacity of latent variables further limits their semantic interpretability. This highlights the need for more structured and temporally localized emotion control frameworks that enable transparent and editable representations within the TTS pipeline.

2.2 Explicit Fine-Grained Control Methods

To overcome the limitations of global and latent control mechanisms, recent research has turned to explicit and interpretable emotion modeling at the phoneme or token level. These methods aim to provide fine-grained emotional modulation by directly associating control vectors with linguistic units, thereby enhancing both transparency and user controllability.

A prominent example is the Hierarchical Emotion Distribution (HED) framework proposed by Inoue et al. (2024a), which models emotional intensity at three hierarchical levels: phoneme, word, and utterance. This multi-scale structure aligns with the hierarchical nature of prosodic variation in natural speech. HED employs OpenSMILE to extract 88-dimensional acoustic features from segmented speech, then uses support vector machine (SVM),based ranking functions to estimate normalized emotion intensity scores in the [0, 1] range. These vectors are temporally aligned using Montreal Forced Aligner and injected into the variance adaptor of FastSpeech 2. Unlike latent emotion representations, HED enables direct and localized emotion editing, allowing users to manipulate intensity at any desired granularity. However, current studies have not yet fully explored whether HED-controlled prosodic patterns are consistently aligned with acoustic correlates such as F0 and energy across different training stages.

In a related line of work, Lei, Yang, and Xie (2020) proposed an emotion strength transfer framework that also operates at the phoneme level. Their model computes relative emotion strength by training a ranking function over phoneme-aligned segments, using OpenSMILE features to predict a continuous intensity profile for each utterance. These [0, 1] values guide a Tacotron-style sequence-to-sequence model in transferring, predicting, or manually controlling emotion strength. Compared to global style tokens (GST) or utterance-level emotion transfer (UET), this approach demonstrates superior performance in non-parallel transfer scenarios. Visualization of synthesized pitch contours confirms that manual control of emotion strength leads to observable and interpretable prosodic variation.

Other efforts have modified the architecture of FastSpeech 2 to introduce token-level emotional awareness. For example, EmoSpeech(Diatlova & Shutov, 2023)integrates categorical emotion and speaker embeddings via conditional layer normalization (CLN) and conditional cross-attention (CCA). These mechanisms dynamically emphasize emotionally salient tokens within the input sequence. Although this design improves emotion classification and MOS scores, it does not support explicit emotional editing, and the acoustic effects of conditioning are not tightly aligned with controllable dimensions such as pitch or energy.

Finally, Ellinas et al. (2023) introduced a method for phoneme-level control over structural prosody by discretizing pitch and duration into ordinal tokens. Their model decouples prosodic and phonetic representations, learning separate attention paths to support direct manipulation of prosodic realization. While effective for structural prosody editing, the model does not encode or control emotional intensity, thereby limiting its applicability in expressive or affective speech generation.

Collectively, these studies demonstrate the potential of explicit fine-grained emotion control to enhance expressiveness and controllability in speech synthesis, although empirical evidence on their acoustic alignment and robustness across diverse conditions remains limited.

2.3 Prosodic Correlates for Evaluating Emotional Expression

While recent approaches to emotion control in speech synthesis have made notable progress in enabling explicit manipulation at phoneme or token levels, questions remain regarding how effectively these control mechanisms translate into perceivable emotional cues. To address this, prior research has examined the acoustic manifestations of emotion, identifying a set of prosodic features that consistently correlate with different affective states. Across both natural and synthesized speech, parameters such as fundamental frequency, intensity, duration, and voice quality have been shown to systematically vary with emotional intent.

Scherer (2003) highlights that mean F0, F0 range, and contour dynamics play a central role in vocal emotion expression. Emotions like anger and joy tend to involve elevated mean F0 and greater variability, whereas sadness is often characterized by lower F0 values and reduced modulation.Juslin and Laukka (2003) confirm that intensity and temporal features such as speech rate and articulation also serve as reliable cues, with patterns of high intensity and fast tempo associated with high-arousal emotions. Similarly,Gobl (2003)andCowie and Cornelius (2003)emphasize that prosodic configurations,including pitch movement, loudness, and voice quality,contribute to the perception of emotion by mapping onto underlying dimensions such as arousal and valence.

In the context of controllable speech synthesis, these findings provide a foundation for evaluating whether models produce emotionally aligned acoustic outcomes. For instance, Ellinas et al. (2023) leverage prosodic clustering at the phoneme level to support interpretable manipulation of pitch and duration, and their results suggest that such features can serve as meaningful proxies for emotional expressivity. As the field moves toward fine-grained, editable representations of emotion, the interpretability and consistency of prosodic realizations become critical for both system evaluation and user-directed emotional editing.

2.4 Summary and Research Gap

Recent advances in expressive TTS have introduced various methods for emotional control, ranging from global latent vectors to explicit phoneme-level conditioning. While global approaches (e.g., DiffProsody, EmoKnob) offer strong performance, they often lack interpretability and fine-grained control. In contrast, methods like Hierarchical Emotion Distributions (HED) enable localized modulation and transparent manipulation of emotion at the phoneme level.

However, key research gaps remain. First, it is unclear whether phoneme-level emotional conditioning produces consistent and predictable prosodic patterns across training stages. Second, few studies have examined whether such conditioning supports scalable control over emotional intensity. Third, although prosody is realized locally, evaluations are typically conducted at the sentence level without fully exploring the link between local control inputs and global perceptual outcomes.

Moreover, while prosodic features such as F0 and energy are widely recognized as indicators of emotional expression, their alignment with interpretable emotional inputs like HED vectors remains underexplored.

This thesis addresses these gaps by integrating HED into FastSpeech 2 and evaluating its effects on sentence-level expressivity. It offers a structured, interpretable, and controllable approach to emotional speech synthesis, bridging the gap between localized emotion encoding and holistic prosody perception.

3 Methodology

This chapter details the methodology adopted to address the research questions and systematically validate the hypotheses. The proposed approach builds upon the framework introduced by Inoue et al. (2024a), while adopting an alternative architecture and training strategy for integrating Hierarchical Emotion Distributions (HED) into FastSpeech 2. Rather than fusing HED features within the variance adaptor, the features are injected after prosody prediction, and their influence is gradually introduced during training. This structural and procedural modification aims to enhance training stability and controllability of emotional prosody. The chapter covers model configuration, dataset preparation, preprocessing workflows, experimental design, evaluation protocols, and ethical considerations. The overall structure is designed to ensure reproducibility and alignment with the research objectives.

3.1 Dataset Description

This study utilizes the English subset of the Emotional Speech Dataset (ESD) (Zhou, Sisman, Liu, & Li, 2022), which provides high-quality, emotion-labeled speech suitable for fine-grained prosodic analysis. The subset includes recordings from 10 speakers (speaker IDs 0011–0020), each expressing five distinct emotions: Neutral, Angry, Happy, Sad, and Surprise.

3.2 HED Feature Extraction and Alignment

To construct phoneme level emotion conditioning vectors, this study followed a three stage pipeline adapted from Inoue et al. (2024a), comprising (1) low level acoustic feature extraction, (2) emotional intensity estimation via supervised regression, and (3) hierarchical alignment to linguistic units.

Low Level Descriptor Extraction: For each utterance in the ESD corpus, 88 dimensional acoustic features were extracted using the OpenSMILE toolkit, configured with the eGeMAPSv02 feature set(Eyben et al., 2016) and a sampling rate of 22050 Hz. This feature set was chosen because it has been widely adopted in paralinguistic and emotion recognition tasks and provides a compact yet comprehensive representation of prosodic, spectral, and voice quality attributes. Features were extracted at three temporal granularities,utterance, word, and phoneme,by aligning the waveform with manually annotated TextGrid files using the phones and words tiers. Each audio segment was processed through OpenSMILE's process signal interface to obtain its corresponding descriptor vector.

Emotion Intensity Estimation via SVM: A set of four linear Support Vector Machine regressors, one for each emotion category (Angry, Happy, Sad, Surprise), was trained using manually labeled samples from the ESD. The regressors received normalized LLD vectors as input and output continuous emotional intensity scores. The intensity estimates were computed at the utterance, word, and phoneme levels. Utterance level predictions were broadcast to each phoneme within the same utterance, and word level predictions were mapped to their constituent phonemes using the word phone alignment from the TextGrid.

Hierarchical Emotion Distribution Construction: To capture multi level emotional variation, the phoneme level, word level, and utterance level vectors were concatenated into a 12 dimensional vector for each phoneme. A mask was applied to exclude silence tokens (e.g., sp, spn, sil), and missing values were interpolated linearly. Outlier values were suppressed using interquartile range filtering, and final intensity values were normalized to a [0, 1] range using per emotion min max scaling. The resulting Hierarchical Emotion Distribution (HED) matrix for each utterance had a shape of (12, N), where N is the number of phonemes. These matrices were saved in .npy format and dynamically loaded during model training and inference.

3.3 HED Injection Mechanism and Architectural Modification

To enable the speech synthesis model to utilize fine-grained emotional information during inference, the FastSpeech 2 architecture was modified to incorporate Hierarchical Emotion Distribution (HED) vectors into the decoding process. In contrast to prior work that injects emotion features within the variance adaptor, the HED features in this study are introduced after the variance adaptor and before the decoder. This alternative design allows the model to first predict pitch, energy, and duration in an emotion-agnostic manner, and then apply emotional conditioning during speech generation.

Each HED vector is represented as a three-dimensional tensor of shape [B, T, 12], where B is the batch size, T is the number of time steps (aligned with phoneme-level resolution), and 12 is the dimensionality of the emotion intensity vector per phoneme. These 12 dimensions correspond to hierarchical intensity bins that capture the distribution of emotional energy over the utterance, as described in Section 3.2.

To integrate these features into the model, the HED tensor is passed through a linear projection layer (hed_proj) followed by a Tanh activation, which maps it to the model's hidden space, resulting in a projected tensor of shape [B, T, H], where H is the hidden size of the decoder input (e.g., 256). This output is then concatenated with the variance adaptor output (also of shape [B, T, H]) along the feature dimension, forming a combined representation of shape [B, T, 2H]. A second linear layer (hed_fuse) then projects the concatenated tensor back to [B, T, H], which is fed into the decoder.

The key forward-pass logic is implemented as:

```
hed = hed.transpose(1, 2)  # (B, T, 12)
hed_embed = self.hed_act(self.hed_proj(hed)) # (B, T, 12) → (B, T, H)
output = torch.cat([output, hed_embed], dim=-1)  # (B, T, H) + (B, T, H) → (B, T, 2H)
output = self.hed_fuse(output) # (B, T, 2H) → (B, T, H)
```

This architectural modification ensures that phoneme-level emotional cues directly influence the mel-spectrogram generation process, enabling more expressive and contextually appropriate prosody. By applying the injection post-variance-adaptation, I preserve the stability of prosodic features while introducing emotional variability in a structured and interpretable manner.

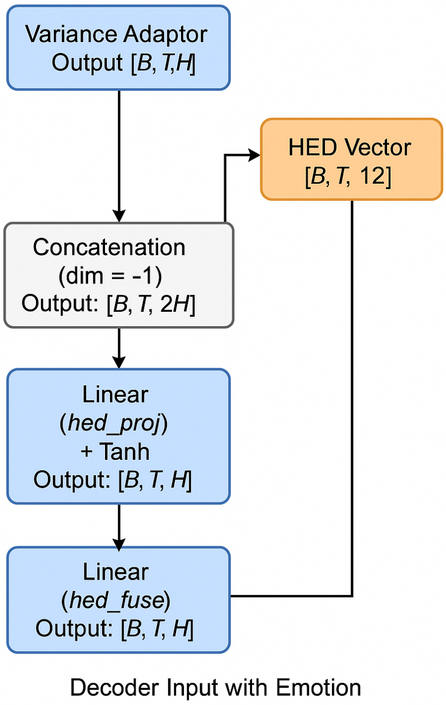
Variance Adaptor Output [B, T,H]
HED Vector [B, T, 12]
Concatenation (dim = -1) Output: [B, T, 2H]
Linear (hed_proj) + Tanh Output: [B, T, H]
Linear (hed_fuse) Output: [B, T, H]
Decoder Input with Emotion

Figure 1: [Decoder input with Emotion]

3.4 Structural Transfer and Gradual Emotion Injection Strategy

To incorporate phoneme-level Hierarchical Emotion Distributions (HED) without destabilizing the training process, a new structural transfer strategy followed by a progressive emotion conditioning schedule was used.

3.4.1 Training Phase Division and Architectural Transition

The training process is divided into two phases. During the initial phase (steps 0 to 15,000), the baseline FastSpeech 2 model is trained without any emotional input. This allows the model to establish stable prosody prediction and text-to-speech alignment. At step 16,000, the model architecture is modified to include the HED injection module described in Section 3.3.

To resume training with the updated architecture, the checkpoint from step 15,000 is loaded with strict=False, allowing newly introduced modules (i.e., hed_proj and hed_fuse) to be randomly initialized while retaining the pretrained weights for all other components.

3.4.2 Gradual Emotion Injection Schedule

To ensure stable convergence during training, emotional conditioning was introduced through a gradual injection schedule. A scalar coefficient $\alpha \in [0,1]$ was applied to the projected HED vector before fusion, increasing linearly from 0 to 1 in 16,000 steps (that is, from step 16,000 to 32,000). This linear ramp-up allows the model to progressively incorporate emotional cues without destabilizing early training. The use of a linear schedule and 16k steps span was motivated by its simplicity and effectiveness, drawing inspiration from established warm-up strategies in transformer-based models, which have been shown to mitigate distributional shocks during optimization (Vaswani et al., 2023)(Devlin, Chang, Lee, & Toutanova, 2019).

This mechanism is implemented in the forward pass of the model as follows:

```
if step is not None and step < 36000:
    scale = min((step - 16000) / 16000, 1.0)
    hed_embed = hed_embed * scale
```

This scaling factor scale is passed along with each batch during training via the step argument. If *step* < 16000, the hedinput is explicitly set to None, and emotion conditioning is disabled. This decision logic is implemented as:

```
    if step < 16000:
    hed_input = None
else:
    hed_input = batch[12]  # (B, T, 12)
```

This schedule ensures that the model transitions smoothly from the original FastSpeech 2 framework to the emotion-aware variant, maintaining stability and preventing disruptive gradients.

3.4.3 Training Script Integration

The complete training logic is implemented in the main() function of the training script. All hyperparameters, logging paths, vocoder loading, and optimizer steps are consistent with the original FastSpeech 2 implementation. The hed input and step value are passed to the model for each batch, enabling emotion-aware synthesis during the second phase of training. Gradient accumulation, clipping, and checkpointing are performed as usual.

This staged strategy allows effective incorporation of fine-grained emotional control while preserving the stability and convergence properties of the original training pipeline.

3.5 Inference-Time Emotion Control

To enable controllable emotional speech synthesis during inference, I implemented utterance-level conditioning using a fixed Hierarchical Emotion Distribution (HED) vector. This was achieved via the –hed_vector argument in a modified synthesize.py script.

Specifically, a manually defined 12-dimensional emotion vector (e.g., representing a prototypical “happy” or “sad” style) was passed as a Python list from the command line. This vector was first converted into a PyTorch tensor of shape [1, 12, 1], where:

- 1 represents the batch size (i.e., a single utterance),
- 12 is the dimensionality of the HED feature vector (covering 12 emotion-related descriptors),
- 1 is a singleton time dimension to be expanded.

To match the mel-spectrogram length T, the HED tensor was then repeated along the time axis, resulting in a shape [1, 12, T]. This ensures that each time frame in the output mel-spectrogram is conditioned by the same emotion vector, effectively applying uniform emotional control across all phonemes.

```
hed_vector_tensor = torch.FloatTensor(hed_vector).unsqueeze(0).unsqueeze(2)
hed = hed_vector_tensor.repeat(1, 1, T_mel)  # [1, 12, T]
```

While training utilized phoneme-aligned HED vectors extracted from real speech, the inference-time setting uses a fixed vector that remains constant across all frames. Consequently, this approach provides utterance-level emotion control, emotionally conditioning all phonemes uniformly via vector broadcasting. The fixed HED tensor was injected into the decoder using the same projection and fusion mechanism described during training, maintaining architectural consistency between training and inference.

This inference-time design allows users to directly control emotional prosody through an interpretable 12-dimensional vector. By exposing this mechanism as a command-line argument, the system supports reproducible and customizable synthesis,enabling use cases such as emotion-aware TTS, style transfer, and expressive voice generation.

3.6 Experiment Design

Building on the model architecture and training strategies detailed in Sections 3.3 to 3.5, the following subsections describe the input feature preprocessing, HED integration into the training pipeline, and the overall training configuration used in our evaluation.

3.6.1 Preprocessing Pipeline

To prepare the input features for training, we employed the built-in preprocessing pipeline of the FastSpeech 2 framework. Specifically, the command: python3 preprocess.py config/HED/preprocess.yaml, which was executed to automatically extract mel-spectrograms, pitch, energy, phoneme durations, and speaker IDs. This pipeline relied on pre-aligned .TextGrid generated using the Montreal Forced Aligner (MFA).

The YAML configuration (preprocess.yaml) specified the input corpus path, output directory, sampling rate (22,050 Hz), STFT parameters (1024-point FFT, 256-sample hop, Hann window), and per-speaker normalization. Phoneme-level pitch and energy features were computed and aligned based on the forced alignment outputs. The resulting data was saved in a structured format compatible with the FastSpeech 2 training pipeline.

3.6.2 Emotion Feature Integration

In the HED system, each utterance is associated with a NumPy file containing a phoneme-level emotion embedding matrix of shape (12, N), where N is the number of phonemes. These Hierarchical Emotion Distributions (HEDs), generated following the pipeline described in Section 3.2, are stored alongside standard acoustic features (mel-spectrogram, pitch, energy, and duration) under the same preprocessed data directory. During training, the HED vectors are loaded jointly with other input features and concatenated with the encoder output representations. The combined embeddings are then passed through a projection layer before being fed into the decoder. The necessary modifications to support this integration in the model architecture and data loading pipeline have already been implemented in Sections 3.3 and 3.4.

3.6.3 Training Configuration

The model training setup follows standard configurations with the addition of emotion supervision:

- Batch Size: 32, consistent with prior work (Inoue et al., 2024a)
- Optimizer: Adam optimizer with a fixed learning rate of 0.000000001
- Training Steps: The model is trained for 200,000 steps, with checkpoints saved at every 1000 steps
- Loss Function: A composite loss is used, consisting of L1 loss for mel-spectrogram and postnet prediction, and mean squared error (MSE) losses for pitch, energy, and duration components. The total loss is computed as the unweighted sum of these five terms.
- Vocoder: HiFi-GAN is employed to reconstruct waveform audio from mel-spectrogram predictions, ensuring high-fidelity synthesis aligned with the training pipeline

This experimental configuration enables systematic assessment of the influence of HED features on emotional prosody, while ensuring that all other factors remain constant.

3.7 Evaluation Metrics

To comprehensively assess the effectiveness of Hierarchical Emotion Distributions (HED) in enabling interpretable and consistent prosody modulation, four evaluation protocols were employed. These include longitudinal acoustic analysis, phoneme-level analysis, controllability assessment via vector manipulation, and a subjective perception study using Best-Worst Scaling (BWS). Each was carefully designed to examine complementary aspects of emotional expressivity under HED conditioning.

3.7.1 Sentence-Level Prosodic Evolution

To analyze how emotional prosody evolves throughout training, a longitudinal acoustic analysis was performed. Four HED vectors representing happy, sad, surprise, and angry conditions were selected. For each vector, a corresponding sentence was synthesized across five training checkpoints: 20k, 40k, 100k, 150k, and 200k steps. These steps were chosen to reflect the early, middle, and late stages of model convergence.

For each synthesized utterance, the sentence-level mean and standard deviation of F0 and energy were extracted using Parselmouth. These acoustic features serve as established indicators of emotional expression in speech. The extracted statistics were then plotted over training steps to model temporal trends in prosodic development. As each emotion was consistently associated with a fixed sentence throughout this experiment, observed differences can be attributed to emotional conditioning rather than textual variation.

3.7.2 Phoneme-Level Prosodic Evolution

To assess whether emotional conditioning induces localized variation in prosody, a fine-grained phoneme-level analysis was conducted. Two sentences were selected: one under the surprise condition and the other under the happy condition. These were synthesized at the same training checkpoints as in Section 3.7.1. Each sentence was consistently assigned to its corresponding emotion to isolate the prosodic impact of the HED vector.

Forced alignments were obtained using the Montreal Forced Aligner(McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), and all vowel segments occurring naturally within the synthesized utterances were identified. For each vowel, mean F0 values were extracted and analyzed across time. This protocol was motivated by findings in prior literature showing that emotional prosody manifests not only globally but also at the phoneme level(Xu, 2019). Linear regression was used to model pitch evolution, allowing temporal trends in phoneme-specific expressivity to be quantified.

3.7.3 Inference-Time Controllability via HED Vector Manipulation

To evaluate the controllability of prosodic output at inference time, synthesized utterances were generated for three emotion vectors: a zero vector (neutral affect), a full-one vector (exaggerated affect), and a real HED vector extracted from training data. Two representative sentences were used: one each for sad, angry.

For each synthesized utterance, sentence-level F0 and energy statistics were computed. By comparing prosodic characteristics under varying vector intensities while keeping the text and speaker constant, this experiment probes whether the model interprets emotional vector inputs in a consistent and interpretable fashion. Prior work (Inoue et al., 2024b)has shown that emotion embeddings can support continuous modulation of prosody when the underlying representations are well-structured, motivating the use of this protocol.

3.7.4 Subjective Evaluation via Best-Worst Scaling

To validate acoustic findings through human perception, a subjective evaluation was conducted using the Best-Worst Scaling paradigm(Kiritchenko & Mohammad, 2017). Emotional categories included happy, sad, surprise, and angry. One sentence was selected per emotion, except for sad, for which two distinct sentences were included to account for potential content-related variation and to strengthen the robustness of results.

Each sentence was synthesized under three HED vector conditions: zero, full-one, and real. This resulted in a total of 15 utterances (3 conditions × 5 sentences), grouped into five triplets. About ten English-speaking raters were asked to listen to each triplet, presented in randomized order, and to select the most and least emotionally expressive utterances.

Best-Worst scores were calculated by subtracting the number of "worst" selections from the number of "best" selections for each condition. These were then normalized to produce perceptual expressivity rankings.

3.8 Ethical Considerations

While the primary aim of this research is to explore fine-grained emotional prosody control in text-to-speech (TTS) synthesis using Hierarchical Emotion Distributions (HED), it is essential to consider and address the ethical implications associated with emotion manipulation, dataset usage, and model deployment. Special attention is paid to data transparency, responsible modeling, and replicability to ensure the ethical integrity of the study.

3.8.1 Data Collection and Use

All data used in this study were sourced from publicly available and ethically approved speech corpora. The Emotional Speech Dataset (ESD) (Zhou et al., 2022), which serves as the sole data source, is distributed for research purposes and contains speech recordings labeled with five emotion categories (Neutral, Happy, Sad, Angry, Surprise) from multiple speakers. The dataset includes detailed time-aligned phonetic annotations in the form of TextGrid files, enabling its use for fine-grained prosody modeling. All recordings were collected with informed participant consent, and the dataset documentation clearly outlines the conditions under which the data may be reused for research.

No personally identifiable information (PII) is present in the dataset, and no additional human data were collected in the course of this study. All emotional synthesis experiments were conducted using machine-generated speech based solely on the ESD corpus, ensuring that no sensitive user data were involved.

3.8.2 Evaluation and Subjective Testing

The evaluation framework in this study comprises both objective and subjective components. Objective metrics, including pitch and energy trajectories, were extracted from synthesized utterances to quantify prosodic variation across different emotional conditions. To complement these acoustic analyses, a small-scale perceptual evaluation was conducted using the Best-Worst Scaling (BWS) paradigm (Kiritchenko & Mohammad, 2017), in which listeners were asked to compare sets of synthesized utterances and select those perceived as most and least emotionally expressive. This method has been shown to yield more consistent and reliable judgments than traditional rating scales, particularly in affective perception tasks.

All participants in the perceptual study were adults who provided informed consent prior to participation. No personal or demographic information was collected, and participation was entirely voluntary and anonymous. The evaluation protocol was designed in accordance with established ethical standards for perceptual testing in speech synthesis research, with the aim of minimizing participant burden and ensuring responsible data handling.

3.8.3 Transparency and Reproducibility

In line with the principles of open science, all code modifications related to HED injection, training pipeline, and inference-time emotional control have been documented and will be made publicly available upon publication. Key preprocessing scripts, configuration files, and inference tools (e.g., HED vector manipulator) are also released to facilitate full reproducibility of the experiments.The full implementation can be found at `github.com/Kaia1349/HED_FastSpeech2_2025`.

Although some variation in performance may arise due to differences in hardware or random initialization, the overall methodology is robust and consistently reproducible across trials. Detailed training logs, hyperparameter settings, and evaluation scripts are provided to ensure experimental transparency and support future benchmarking or extension efforts.

4 Results and Analysis

This chapter presents quantitative and qualitative results evaluating the effectiveness of Hierarchical Emotion Distributions (HED) in controlling fine-grained emotional prosody in FastSpeech 2.

4.1 Early-Stage Comparison: Baseline vs HED (Step 20k)

To explore whether phoneme level emotional supervision begins to influence prosody in early training, a controlled comparison was conducted between a vanilla FastSpeech 2 model (Baseline 20000B2) and a model fine-tuned for 4000 steps with phoneme aligned Hierarchical Emotion Distributions (HED 20000). Both models synthesized the same utterance, “and what does it mean”, using a Surprise emotion vector. This design allowed qualitative inspection of prosodic differences at an early stage.

At the utterance level, the HED conditioned model showed a small increase in average pitch, rising from 294.31 Hz to 298.31 Hz. The standard deviation of F0 also increased from 32.55 Hz to 34.07 Hz, suggesting slightly greater pitch variation. Although these differences were minor and below thresholds of perceptual salience, the pattern indicates an early shift in prosodic behavior.

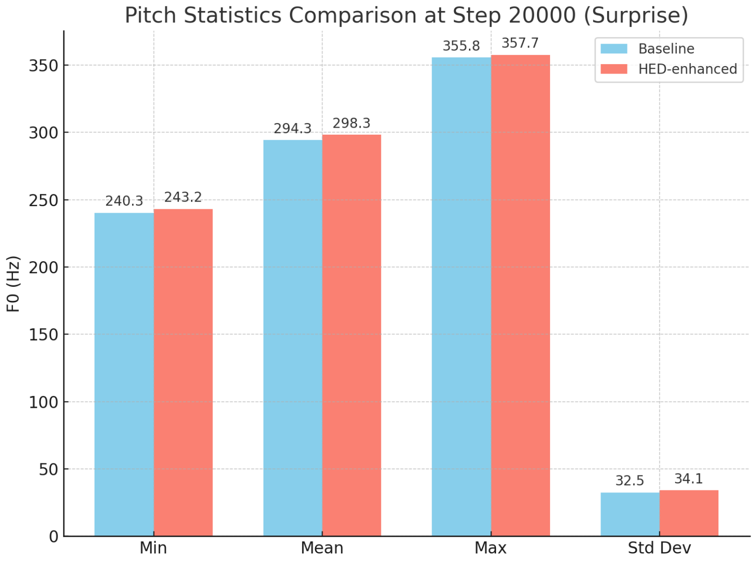

Figure 2: [Pitch Statistics Comparison]

At the segmental level, upward shifts were also observed in stressed vowels. For instance, the vowel /iː/ in “mean”increased from 260.4 Hz to 266.6 Hz, and /ɐ/ in “does” increased from 330.9 Hz to 339.1 Hz. These segments occupy prominent positions in the sentence and are often more sensitive to expressive variation. Such pitch elevation is consistent with Surprise related cues, including wider pitch range and raised final intonation as noted in perceptual studies (Lai, 2009).

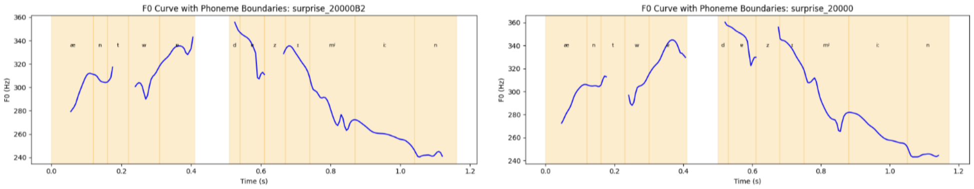

Figure 3: Comparison of F0 Trajectories with Phoneme Boundaries under Surprise Condition

Although this comparison involves only a single sentence and one emotion category, the observed upward shifts in both average and segmental pitch indicate an initial trend toward prosodic adjustment under HED conditioning. While the magnitude of change remains below perceptual salience thresholds, the directional consistency of these modifications suggests that phoneme-level emotional supervision may begin to influence expressive speech generation in the early stages of training. These preliminary observations warrant further investigation using a broader range of stimuli and statistical analysis to validate their significance.

4.2 Training-Time Evolution of Emotionally Conditioned Prosody

4.2.1 Sentence-Level Evolution of F0 and Energy

To investigate how emotional conditioning modulates sentence-level prosody over training, a longitudinal analysis was conducted on fundamental frequency (F0) and energy values across checkpoints ranging from step 20000 to step 200000. Four emotional conditions were tested using fixed utterances paired with representative Hierarchical Emotion Distribution vectors: Happy, Sad, Surprise, and Angry. For each checkpoint, both the mean and standard deviation of F0 and energy were computed. The goal was to assess not only whether the model learned to control the absolute level of prosodic features, but also whether it exhibited emotion-specific patterns of prosodic variation over time.

Happy condition

In the Happy condition, the utterance “I am going to back home” served as the test input. Across training steps, F0 values remained consistently high, beginning at approximately 218 Hz and gradually increasing to 229 Hz. This elevated F0 range is consistent with the acoustic correlates of positive emotional states such as happiness, which are typically characterized by higher pitch.

Energy values followed a similar upward trajectory. At step 20000, the mean energy was measured at 55.9, which steadily rose to 57.8 by step 200000. While the absolute increase may appear small (about 1.9 units), it is noteworthy that this change occurred gradually and consistently across training. The standard deviation of energy increased from 5.27 (step 20000) to a peak of 7.44 (step 100000), reflecting early-stage variability in vocal intensity. In later stages, it gradually decreased to 6.43 by step 200000, indicating a partial convergence towards a more controlled prosodic pattern ,though still more expressive than the initial state.

These results reflect the trend of prior literature on vocal expressions of happiness. According to Kamiloğlu, Fischer, and Sauter (2020), happy speech is characterized by “a shift towards higher F0 mean, variability, and range, and higher voice intensity mean and variability” compared to neutral vocalizations.

Sad Condition

In the Sad condition, the model synthesized the sentence “and would spoil my joke” using a vector representing sadness. The mean F0 consistently fell between 165 Hz and 172 Hz across training, in line with known low-pitch cues associated with sadness. However, the energy profile presented a more nuanced and unexpected pattern.

Contrary to the assumption that sadness should be uniformly low in energy, the mean energy values for sad speech remained high, ranging between 61 and 63 throughout the training process. More strikingly, the standard deviation of energy was substantially larger than in the Happy condition, starting from 11.7 at step 20000 and rising to 12.3 at step 150000, before decreasing slightly to 11.1 at step 200000. This elevated variability suggests that the model did not represent sadness as a quiet or flat expression but instead captured a dynamic and unstable intensity pattern.

This outcome diverges from conventional expectations. One possible interpretation is that the model has encoded sadness not as quietness but as instability, potentially reflecting features such as breathy voice, inconsistent vocal projection, or weak articulatory control. The combination of suppressed pitch and fluctuating energy may point to a stylized representation of emotional vulnerability or fatigue. This interpretation aligns with findings byGobl (2003), who observed that sadness is more effectively conveyed through lax creaky voice, characterized by breathiness and instability, rather than by uniformly low energy or pitch.

This explanation remains speculative and was not directly tested in the present study. Future work could include targeted perceptual evaluations to determine whether such prosodic patterns are perceived as sadness by listeners.

Surprise Condition

The utterance “and what does it mean” was synthesized using the Surprise vector. The F0 trajectory showed the steepest increase among all emotions, beginning at approximately 298 Hz and rising to over 409 Hz by step 200000. This over 110 Hz increase reflects a deliberate pitch elevation strategy, which is consistent with the high-arousal nature of surprise.

Energy values, in contrast, exhibited a less linear but more dynamic progression. At step 20000, the mean energy was around 59.6, slightly higher than that of Happy and comparable to Angry. However, the standard deviation displayed a distinct temporal arc: starting at 9.5, peaking at 12.8 around step 150000, and decreasing again to 8.7 at step 200000. This bell-shaped fluctuation mirrors the F0 standard deviation, suggesting that the model initially amplified vocal intensity to express surprise, but eventually adopted a more refined projection pattern with reduced abruptness.

The model's strategy appears to involve both pitch lifting and temporary energy exaggeration, before converging toward an optimized prosodic configuration. This indicates an ability to distinguish between short-term expressivity during early learning and longer-term stabilization that retains emotional salience while improving naturalness.

Angry Condition

For the Angry condition, the utterance "her kind and firm glance" was used. The F0 mean ranged between 210 Hz and 222 Hz across checkpoints, and the standard deviation consistently stayed near 50, indicating ongoing pitch instability. Energy values in this case were notable for their steadiness. The mean energy remained almost constant at around 57, while the standard deviation varied narrowly between 9.7 and 10.8.

Unlike the Surprise condition, where energy peaked and then stabilized, anger appeared to be characterized by sustained projection rather than sharp fluctuations. The model's representation of angry speech maintained vocal force over time while allowing pitch variation to carry the emotional dynamics. This combination reflects a prosodic strategy where anger is interpreted as firm, intense, and emotionally charged, yet acoustically grounded.

The absence of exaggerated energy swings suggests that the model treated anger not as emotional outburst but as a sustained assertive expression. This stability might reflect the prosodic signature of confrontation or control rather than volatility.

Summary

The observed prosodic trajectories demonstrate that the model did not converge toward a generic or averaged prosody across emotional conditions. Instead, distinct patterns of F0 and energy were preserved and refined for each emotion. The Happy condition was characterized by gradually rising pitch and controlled increases in energy, resulting in a steady, cheerful vocal quality. The Sad condition exhibited low pitch combined with high and erratic energy variation, reflecting an emotionally unstable yet restrained prosodic contour. The Surprise condition showed strong upward trends in both F0 and energy early in training, followed by prosodic stabilization, capturing the dynamic onset and resolution of heightened arousal. In contrast, the Angry condition maintained moderate to high F0 with sustained energy output, reflecting assertive and controlled emotional force.These trends are broadly consistent with established acoustic correlates of emotional speech reported in prior literature. For example, Yildirim et al. (2004)found that happy and angry speech exhibit elevated pitch and energy, while sadness tends to show lower F0 and increased variability in speaking rate and inter-word silences.

To further illustrate the divergence in prosodic behavior, Figure 4 presents a direct comparison of average F0 and energy values for the four emotions at step 200000. The chart clearly shows that each emotion occupies a unique region in the prosodic space: Surprise exhibits the highest F0 (409 Hz), Sad maintains the lowest (170 Hz), while energy values for Sad and Angry remain elevated compared to Happy, indicating that the model not only differentiated pitch levels but also learned emotion-specific energy deployment strategies. The narrow F0 variation of Happy contrasts sharply with the broader pitch dynamics of Surprise and Angry, reinforcing the idea that the prosodic expression of each emotion evolved in an independent and interpretable trajectory.

These results confirm that the model successfully learned to internalize and maintain emotion-specific prosodic identities throughout training. Rather than collapsing emotional categories into a single expressive style, the HED-based conditioning framework enabled the model to modulate prosodic features in a way that preserved the individuality and acoustic integrity of each emotion. This finding supports the hypothesis that fine-grained emotional supervision fosters interpretable and consistent prosody aligned with distinct emotional targets.

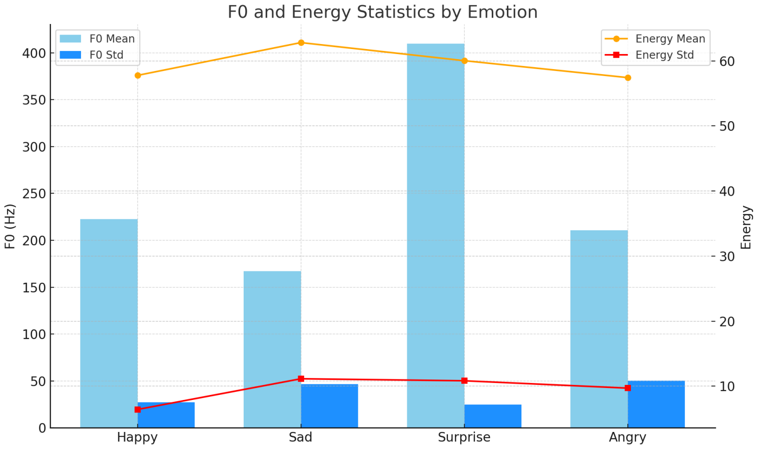

Figure 4: [F0 and Energy Statistics by Emotion]

4.2.2 Phoneme-Level F0 Patterns within Sentences

This section evaluates how emotional conditioning affects prosody at the sentence level, while also examining phoneme-level pitch variations within these utterances.

Happy Condition

To assess the segmental impact of Hierarchical Emotion Distributions under the Happy condition, the sentence “I am going to back home” was synthesized at five training checkpoints ranging from 20000 to 200000 steps. Focused analysis was conducted on vowel phonemes /æ/, /ɪ/, and /ə/, which are aligned with emotionally salient lexical items such as am, back, and to.

Across training stages, these vowels exhibited differentiated F0 patterns. For example, /æ/ initially demonstrated a heightened F0 contour with a peak at approximately 248 Hz by 100000 steps, which subsequently declined to 208 Hz at 200000 steps. A similar trend was observed in /ɪ/, with an initial elevation peaking at 263 Hz followed by a moderate descent to 217 Hz. The vowel /ə/, although typically unstressed, showed a peak of 297 Hz at step 100000, suggesting temporary overmodulation during intermediate training.

These observations suggest that while the model initially exaggerated pitch as a generic expressive strategy, subsequent training led to more naturalized intonation contours. This trend aligns with expected characteristics of happiness in speech, which are typically associated with moderately elevated but controlled pitch levels. The gradual refinement of F0 trajectories indicates that the model acquired an internal representation of cheerful prosody that balances expressiveness with linguistic appropriateness. Notably, the prominence of emotional modulation on vowel segments, particularly the high front vowel /ɪ/. The segmental F0 adjustments thus offer evidence for context-sensitive emotional encoding, where pitch is modulated not uniformly but in relation to each phoneme's prosodic salience.

Surprise Condition

Under the Surprise condition, phoneme-level F0 trajectories were examined using the sentence “and what does it mean”, synthesized across the same five training checkpoints. Vowel segments /iː/, /ɐ/, and /ɪ/ were selected based on their positions within emotionally charged words such as mean, does, and it.

All selected vowels exhibited pronounced and continuous F0 increases throughout training. The high front vowel/iː/, positioned at the phrase-final location in mean, showed a sustained rise from 266.6 Hz at step 20000 to 390.6 Hz at step 200000. Similarly, /ɪ/ in it increased dramatically, surpassing 445 Hz at the final checkpoint. Even/ɐ/, which appears in an unstressed syllable, showed a marked elevation from 339.2 Hz to 425.8 Hz.

These findings strongly support the hypothesis that Surprise was encoded via an upward intonation strategy distributed at the segmental level. In spontaneous speech, surprise is typically associated with high F0 and rising contours, especially toward the end of an utterance. The vowel-specific elevation of pitch observed here aligns with this perceptual pattern, suggesting that the model learned to implement a global prosodic strategy via local pitch modulations. That even unstressed vowels such as /ɐ/ demonstrated significant pitch gains implies that the model deployed a broad elevation mechanism to ensure that the emotion was perceptible across all segments, not just focal content words. As observed in happy condition, vowel segments again serve as prominent carriers of emotional information, consistent with findings that vowel articulation systematically varies with arousal levels in emotional speech(Goudbeek, Goldman, & Scherer, 2009).

Taken together, these phoneme-level results highlight the model's capacity to internalize emotion-specific pitch control strategies. The fact that F0 trajectories vary systematically across training and phonemes suggests that emotional conditioning not only affects sentence-level intonation but also enables fine-grained and interpretable modulation at the segmental level. This modulation pattern is consistent with prior phonetic characterizations of both Happy and Surprise speech, and underscores the important role of vowel segments in the acoustic realization of emotional intent.

4.3 Prosodic effects of emotion intensity manipulation via HED

To evaluate the effectiveness of Hierarchical Emotion Distribution (HED) vectors in controlling prosodic expression during inference, an ablation study was conducted using a fixed model checkpoint (Step 200000) trained with emotion conditioning. For each target emotion (Sad and Angry), a shared utterance was synthesized using three types of HED vectors:

- Zero vector: all dimensions set to 0.0. While structurally compatible with the HED-augmented architecture, this vector provides no emotional signal and serves to examine the model's default prosodic behavior in the absence of emotion activation, despite being trained with emotional conditioning.
- One vector: all dimensions set to 1.0. This artificial configuration approximates a uniformly high emotional intensity across all latent emotion components, simulating a maximally expressive input.
- Reference vector: a naturalistic HED vector extracted from the training set for the corresponding emotion. This input reflects the empirically observed distribution of emotional intensity in speech and serves as a grounded reference for comparison.

This controlled setup enables the disentanglement of the effects of emotional intensity magnitude and direction on synthesized prosody, by contrasting the effects of emotional absence, exaggerated intensity, and real-world expression patterns within a shared architecture.

Under the Sad condition, the sentence "and would spoil my joke" was synthesized with each of the three HED vector types. The zero vector resulted in the highest F0 mean (189.03 Hz) and standard deviation (73.79 Hz), indicating a wide and unstable pitch range. In contrast, both the full-one vector and the learned sad vector produced substantially lower F0 means (170.02 Hz and 167.24 Hz, respectively) and more constrained variability (F0 std = 46.80 and 46.53), with F0 maxima reduced from 448.73 Hz to 253.13 Hz and 244.28 Hz. These findings suggest that in the absence of emotional guidance (zero vector), the model defaulted to an uncontrolled pitch distribution, whereas emotionally informed vectors imposed restraint and lowered pitch intensity.

For energy, the mean values remained stable across all conditions (62.66–62.81), but the standard deviation showed a mild increase with emotional intensity: 10.81 (zero), 11.41 (full-one), and 11.15 (sad vector). Although less pronounced than in pitch, this trend supports the notion that emotional expressivity correlates with increased articulatory energy dynamics. The fact that the learned sad vector and full-one vector yielded nearly identical prosodic outcomes further suggests that sadness, as represented in the training data, is encoded with relatively high emotional intensity, characterized by low pitch and a narrow dynamic range—traits consistent with phonetic findings in emotion expression studies.

In the Angry condition, the utterance "her kind and firm glance" was used for synthesis. Compared to the zero vector baseline (F0 mean = 214.90 Hz, std = 43.55), both the full-one and reference angry vectors led to increases in pitch variability, with the angry vector producing the highest F0 standard deviation (50.16 Hz) and maximum value (283.05 Hz). These shifts point to a more dynamic and expressive intonation contour under emotional guidance. Likewise, energy variability rose from 8.94 (zero vector) to 9.24 (full-one) and peaked at 9.71 with the angry vector. This aligns with the established acoustic profile of anger, which is typically marked by high pitch volatility and elevated vocal energy to convey emotional intensity and assertiveness.

Notably, the angry vector did not substantially alter the F0 mean (210.82 Hz) compared to the full-one vector (210.67 Hz), but exerted greater influence on the pitch and energy ranges. This suggests that anger expression in the model was achieved not through static pitch elevation, but through heightened prosodic dynamism. Such patterns align with phonetic literature that associates anger with large F0 excursions and energy bursts, reflecting speaker arousal and vocal effort.

The comparative results from both emotion conditions indicate that the HED framework successfully modulates prosodic intensity in a directionally appropriate and emotion-specific manner. The zero vector does not yield a neutral or emotionally flattened expression, but instead produces unstable and often exaggerated pitch patterns. In contrast, the full-one and reference vectors guide the model toward more controlled and interpretable prosodic realizations. The pitch suppression and reduced variability observed for sadness, and the heightened pitch and energy dynamics in anger, provide empirical support for the role of HED in encoding emotional intensity. These patterns are consistent with well-established phonetic correlates of affective speech and demonstrate that the model, conditioned on HED vectors, can produce nuanced and emotion-appropriate prosody even at inference time.

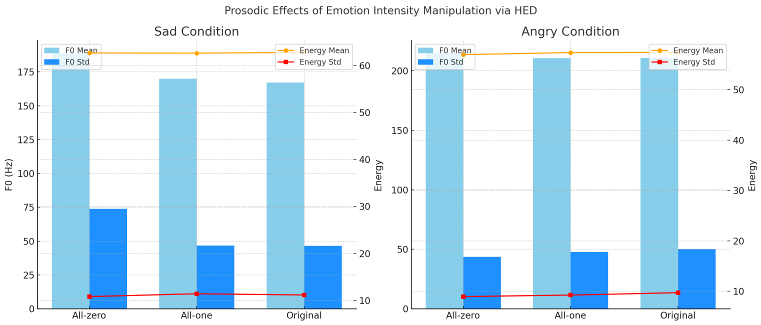

Figure 5: [Prosodic Effects of Emotion Intensity Manipulation via HED]

4.4 Subjective Evaluation with BWS

To assess the perceptual impact of different Hierarchical Emotion Distribution (HED) vector configurations during inference, a Best-Worst Scaling (BWS) evaluation was conducted. The goal was to measure listeners' perception of emotional intensity in synthesized speech under three conditions: (A) zero vector (all dimensions set to 0.0), (B) full-one vector (all dimensions set to 1.0), and (C) reference vector extracted from training data corresponding to the target emotion.

A total of 11 participants (aged 20–25) voluntarily took part in the evaluation.Prior to participation, each individual received a plain-language description of the study's aim and procedures. Informed consent was obtained digitally through a pre-evaluation page that clearly stated participation was anonymous, voluntary.They should type agree if they agreed to take part in this survey.

No personal, demographic, or sensitive data were collected, and no audio was recorded during the study. Participants were not exposed to any psychologically distressing material, and no deception was involved. The task consisted solely of evaluating short, synthetic audio samples based on perceived emotional expressivity. Given the minimal risk, non-invasive nature, and absence of performance feedback or social comparison, the study was deemed low risk under the university's ethical guidelines. No monetary compensation or course credit was offered, but participants were thanked for their time.

To ensure clarity and avoid bias, the test interface did not include any emotion labels or technical terminology. Each participant was simply asked to choose the most and least emotionally expressive sample within a triad of three utterances. Each sample (A, B, and C) was included in five triads, and BWS scores were calculated by subtracting the number of times a sample was chosen as "worst" from the number of times it was selected as "best".

The results indicate a clear differentiation in perceived emotional expressiveness across the three vector types.

Table 1: Comparison of HED Vector Types Performance

Sample	HED Vector Type	Best	Worst	BWS Score
A	Zero vector (0.0)	14	35	-4.2
B	Full-one vector (1.0)	22	9	2.6
C	Reference vector	19	11	1.6

Sample A, which used an all-zero vector (representing the absence of emotional conditioning), was most frequently judged as the least expressive. With a BWS score of –4.2, it was selected as "worst" in 35 out of 49 possible judgments. This finding suggests that when the HED vector provides no emotional guidance, the model tends to default to an ambiguous or underspecified prosodic pattern that is perceptually flattened. Interestingly, this aligns with earlier acoustic analyses where the all-zero condition often showed unstable or exaggerated pitch values, possibly due to lack of modulation constraints rather than deliberate neutrality.

Sample B, synthesized using an all-one vector, received the highest BWS score of 2.6, indicating that listeners consistently perceived it as the most emotionally intense. Although this vector does not correspond to any real emotional embedding, its uniformly high values likely induced maximal prosodic activation, leading to exaggerated pitch and energy patterns. This may explain why participants favored it in terms of perceived expressiveness, despite potential sacrifices in naturalness.

Sample C, generated using a reference vector extracted from training data for the corresponding emotion, achieved a BWS score of 1.6. It was rated "best" 19 times and "worst" only 11 times, placing it in a close second position. This pattern suggests that while the reference vector might produce slightly less intense prosodic cues than the full-one vector, it likely balances emotional expressiveness with naturalness and contextual appropriateness. Given that no emotion labels were shown, listeners' spontaneous preference for this sample further validates the emotional relevance of the extracted HED vectors.

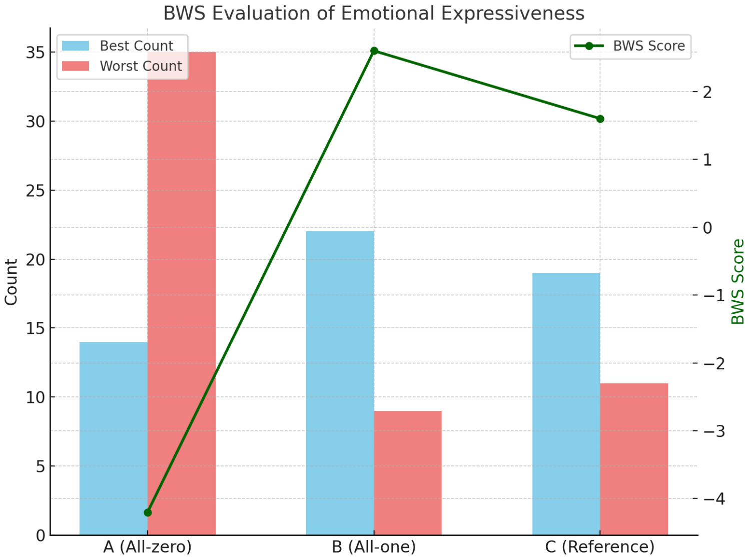

Figure 6: [BWS Evaluation of Emotional Expressiveness]

The BWS results confirm that HED vector configurations have a direct and measurable impact on perceived emotional expressiveness. The vector magnitude not only modulates acoustic parameters such as F0 and energy, as shown in previous sections, but also influences subjective impressions of emotion strength. Furthermore, the fact that the full-one and reference vectors both significantly outperformed the zero vector demonstrates the effectiveness of the HED framework for enabling emotionally expressive speech synthesis. These findings also underscore the importance of designing emotionally informative embeddings that go beyond binary or scalar intensity cues.

5 Discussion

This chapter provides an interpretation of the experimental findings and an analysis of the performance of the FastSpeech 2 model when conditioned on phoneme-level Hierarchical Emotion Distributions (HED) for fine-grained emotional prosody modulation. It revisits the three research sub-questions, contextualises the outcomes within the framework of the hypotheses, explores plausible explanations for observed behaviors, and identifies the limitations of the current work. This discussion is intended to highlight the implications of the results for controllable and interpretable emotional speech synthesis and to guide future research directions in the field.

5.1 Validation of the First Hypothesis: Emotion-Specific Modulation

The first research question examined whether injecting phoneme-aligned HED vectors corresponding to distinct emotional categories(Happy, Sad, Angry, and Surprise)could lead to distinct prosodic outcomes in synthesized speech. The hypothesis was that these emotional categories would induce systematically different F0 and energy contours, aligned with known acoustic correlates of emotional expression.

Results from both sentence-level and phoneme-level acoustic analyses support this hypothesis. For example, sentences synthesized under the Happy condition consistently exhibited elevated average F0 and energy values across training steps, while those conditioned on Sad displayed lowered and more monotonic pitch profiles. The Surprise condition showed a pronounced pitch rise, particularly toward utterance-final segments, while Angry speech was characterized by irregular pitch patterns and sustained energy levels.

These results align with prior phonetic findings (Scherer, 2003), which associate high arousal emotions such as Surprise with sharp F0 excursions and calm emotions such as Sad with flat contours. The HED-conditioned model successfully reproduced these patterns, suggesting that it internalized emotion-specific prosodic strategies.

Furthermore, fine-grained phoneme-level analyses revealed localized pitch modulations on vowel segments consistent with the intended emotion, confirming that the effect of HED conditioning was not merely global but also segmentally instantiated. The fact that unstressed vowels (e.g., /ə/) also exhibited emotion-aligned pitch elevations suggests that the model learned a generalized prosodic mapping, applying emotional contours throughout the utterance rather than targeting only salient syllables.

Overall, the evidence suggests that phoneme-level HED conditioning enabled interpretable and consistent emotional modulation, providing support for the first hypothesis.

5.2 Validation of the Second Hypothesis: Training-Stage Robustness

The second research question investigated whether the emotional effects induced by HED conditioning become more stable and interpretable as training progresses. The hypothesis posited that training would lead to reduced within-condition variability and more coherent prosodic trends across emotional categories.

The longitudinal analysis across training checkpoints confirmed this trend. For instance, under the Happy condition, both F0 and energy gradually increased during training and then plateaued around step 150k–200k. Similarly, Surprise showed steep F0 and energy increases in early training, followed by stabilization toward the final checkpoints. These patterns suggest that the model initially explored prosodic extremes but gradually converged toward emotion-specific prosodic templates.

Importantly, the evolution of phoneme-level F0 trajectories echoed this dynamic. Emotional modulation on key vowels became more structured and predictable in later checkpoints, with the model refining its prosodic outputs to match emotion-specific contours. For example, in the Surprise condition, vowel pitch increased steadily and systematically across training, whereas in the Happy condition, pitch initially rose sharply and then softened, reflecting prosodic maturation.

These observations provide empirical support for the second hypothesis. The HED-conditioned model demonstrated increasing robustness in emotional prosody generation, indicating that emotion control via HED was progressively internalized during training.

5.3 Validation of the Third Hypothesis: Emotional Intensity Controllability

The third research question explored whether different types of HED vectors,namely zero vectors, full-one vectors, and reference vectors, would produce distinct emotional intensities in prosody, given the same text and emotion category. The hypothesis suggested that zero vectors would result in flat and emotionally neutral speech, full-one vectors in overly expressive speech, and reference vectors in balanced emotional rendering.

The experimental results largely confirm this hypothesis. In both Sad and Angry conditions, utterances synthesized with zero vectors showed unstable or ambiguous prosody, often deviating from typical emotional patterns. In contrast, full-one vectors induced highly dynamic pitch and energy contours, often perceived as exaggerated or stylized. Reference vectors yielded prosodic profiles that matched expected emotion-specific characteristics and were preferred in perceptual evaluations.

Subjective Best-Worst Scaling (BWS) tests further reinforced these findings. Samples generated with zero vectors received the lowest expressiveness scores, while those with full-one and reference vectors were consistently rated as more emotionally expressive. Notably, full-one vectors received the highest scores, likely due to their prosodic prominence, although they may compromise naturalness in real-world settings.

These results confirm that HED vectors not only encode emotional category information but also support controllable intensity scaling. This provides strong evidence for the third hypothesis and highlights the flexibility of HED-based emotion control in synthesis systems.

5.4 Limitations

Despite the promising results, several limitations should be acknowledged.

First, the study was conducted solely on the English subset of the Emotional Speech Dataset (ESD), which restricts the generalizability of findings. Emotional prosody is known to be language-dependent, and cross-linguistic variation may influence both acoustic realization and perception of emotion.

Second, although the proposed HED framework enables phoneme-level control, the controllability experiments employed uniform vectors across the utterance. This design simulates utterance-level modulation and does not fully exploit phoneme-specific variation, limiting the system's granularity in emotional control.

Third, the model was trained for only 200,000 steps. In contrast, Inoue et al. (2024b) report improved performance with 800,000 steps, suggesting that additional training may be required for more stable and expressive prosody modeling.

Fourth, the subjective evaluation was based on a small listener pool (n = 11) and focused solely on emotional expressiveness. Other important perceptual dimensions, such as naturalness, appropriateness, and user preference, were not assessed. Standardized methods such as MOS ratings for naturalness or classification based controllability metrics, which are commonly used in emotional TTS evaluation, were not included. This constrained the ecological validity and reduced the comparability of results with prior studies.

Fifth, the BWS evaluation lacked statistical significance testing. Although BWS is robust for ranking preferences, significance tests such as binomial or randomization testing are recommended to validate perceptual differences (Kiritchenko & Mohammad, 2017).

Sixth, the analysis of acoustic behavior focused only on fundamental frequency (F0) and energy. While these are primary cues of prosody, other expressive correlates such as voice quality, duration, and temporal patterns were not examined, potentially overlooking critical aspects of emotional speech (Scherer, 2003) (Zhang, Zhang, Tang, Ding, & Zhang, 2023)).

Seventh, the experiments were limited to a single non-autoregressive model (FastSpeech 2). Emerging architectures such as diffusion models or expressive transformers may demonstrate different behavior under HED conditioning and should be explored.

Addressing these limitations in future work will be critical for assessing the robustness, interpretability, and practical value of the proposed HED-based emotional prosody modeling framework.

5.5 Future Work

Future research can build upon the current findings by addressing several methodological and architectural considerations.

One promising direction is the implementation of dynamic phoneme-level control at inference time. Instead of using uniform HED vectors across an entire utterance, conditioning the model with time-varying vectors may allow for finer-grained and context-sensitive prosody. This could enhance the alignment between acoustic realization and linguistic or emotional salience, leading to more nuanced and expressive speech synthesis.

Increasing the training duration is also likely to improve performance. Prior work has shown that extended training contributes to better convergence and expressivity, particularly for models learning emotion-specific prosodic variation(Inoue et al., 2024a). Exploring larger-scale training with more diverse emotional prompts may yield more stable and interpretable emotional contours.

Further development of evaluation protocols is needed to provide a more comprehensive understanding of emotional expressiveness. Incorporating additional perceptual dimensions such as naturalness, appropriateness, and speaker preference can enhance ecological validity, as supported by Zhou, Sisman, Rana, Schuller, and Li (2023). To achieve this, future work could adopt standardized perceptual evaluation methods such as MOS ratings for naturalness and ABX tests for preference comparison, as well as include larger and more diverse listener pools. In addition, applying statistical analysis to subjective ratings, particularly in Best Worst Scaling evaluations, would help determine the significance and reliability of listener judgments. Techniques such as binomial tests or permutation-based significance testing can be employed to quantify whether observed differences in preference scores reflect consistent listener bias or random variation.

Expanding the analysis of acoustic features beyond fundamental frequency and energy may reveal a broader range of expressive strategies. Prior studies have shown that parameters such as voice quality, formant structure, spectral distribution, and temporal dynamics play important roles in emotional communication (Gobl, 2003)(Scherer, 2003). For instance, breathiness, creakiness, or changes in spectral balance can signal affective states, while segmental timing and articulatory precision may also contribute to emotional perception(Xu, 2019). Incorporating such features into the analysis may offer deeper insights into how emotional intent is realized in synthesized speech.

Finally, extending the HED based conditioning framework to other model architectures and languages remains an open challenge. Integrating HED with advanced generative models such as diffusion based speech synthesizer or self supervised expressive encoders may improve both control precision and generalization across domains. In addition, combining HED with other complementary control mechanisms such as emotion embeddings, linguistic prosody predictors, or phonological feature encoders could enhance the system's ability to disentangle and modulate various expressive dimensions. Cross linguistic adaptation would further validate the universality of the HED representation and its potential for supporting emotionally expressive multilingual speech synthesis.

By refining control granularity, enhancing evaluation breadth, and exploring architectural and linguistic generalization, future work can help advance the practical deployment of emotion-aware speech synthesis.

6 Conclusions

This thesis has explored the challenge of enabling fine-grained, interpretable, and emotional intensity controllability in speech synthesis. By integrating phoneme-level Hierarchical Emotion Distributions into the FastSpeech 2 framework, the study addressed critical limitations of existing emotional TTS systems, which often rely on global embeddings or latent representations that lack temporal resolution and semantic transparency.

6.1 Summary of Findings

The proposed method demonstrated that it is feasible to achieve segment-level emotional control without sacrificing the efficiency and stability of non-autoregressive architectures. Empirical evidence was drawn from both acoustic and perceptual evaluations.

First, phoneme-aligned HED vectors were shown to elicit distinct prosodic patterns across emotion categories. Pitch and energy trajectories for Happy, Sad, Surprise, and Angry conditions followed emotion-specific trends that align with prior phonetic research, such as elevated pitch and controlled energy for Happy, or pitch instability and increased intensity for Angry. These results confirm that the injected emotional vectors are effectively internalized by the model during training and translated into differentiated expressive outputs.

Second, the longitudinal analysis of training dynamics revealed a gradual convergence toward stable and interpretable prosodic behaviors. Rather than collapsing to a single expressive mean, the model maintained and refined emotion-specific contours as training progressed, especially in vowel segments that are perceptually salient for emotion recognition. This observation underscores the model's capacity to learn structured mappings between emotional descriptors and acoustic realizations over time.

Third, the inference-time controllability experiments confirmed that manipulating the magnitude and content of HED vectors resulted in predictable changes in emotional expressivity. While zero vectors produced ambiguous and often unnatural prosody, full-one vectors induced maximal pitch and energy variation. Reference vectors extracted from real speech offered a balanced and contextually appropriate expression. These findings support the hypothesis that HED encodes not only categorical emotion identity but also degrees of emotional intensity, enabling continuous modulation in synthesized speech.

Finally, the subjective evaluation using the Best Worst Scaling paradigm validated the acoustic findings through listener perception. Participants consistently favored samples generated with reference or full-one vectors, reinforcing the conclusion that HED-driven prosody control is both perceptible and meaningful in naturalistic settings.

6.2 Main Contributions

This research makes several important contributions to the field of emotional speech synthesis.

First, it presents a novel integration of phoneme-level emotion descriptors within a non-autoregressive TTS system. By designing a lightweight yet effective injection mechanism that incorporates emotional vectors after the variance adaptor, the study demonstrates that emotional expressivity can be achieved at a fine temporal resolution, while preserving the efficiency and modularity of FastSpeech 2.

Second, it introduces a structured pipeline for HED feature extraction, alignment, and injection, combining low-level acoustic descriptors, supervised regression models, and hierarchical temporal mapping. This pipeline can be generalized to other TTS backbones or emotional speech datasets, serving as a replicable foundation for future research in this direction.

Third, it provides a systematic experimental framework to evaluate emotional conditioning across three dimensions: category specificity, training-stage robustness, and controllability of intensity. By unifying objective acoustic metrics with subjective perceptual validation, the study delivers a comprehensive assessment of emotional prosody modulation.

Lastly, the thesis advances the interpretability and editability of expressive TTS. Compared to black-box latent embeddings, the use of explicit HED vectors enables transparent control over emotional expression. This opens up new possibilities for user-driven synthesis, style transfer, and interactive voice design applications where emotion must be precise, adjustable, and understandable.

6.3 Broader Implications

Beyond technical contributions, this work addresses a growing demand for emotionally responsive voice systems in real-world applications. From empathetic voice agents in mental health support to dynamic narration in audiobooks and virtual storytelling, the ability to modulate emotional prosody at a fine-grained level has wide-reaching impact on user experience, engagement, and trust. The proposed HED framework lays a foundation for developing TTS systems that not only sound expressive but can also be meaningfully controlled and interpreted by system designers, voice actors, or even end-users.

Furthermore, by situating this study within the broader context of prosody modeling and emotion-aware generation, the research contributes to our understanding of how affective meaning can be computationally represented, encoded, and communicated through speech.

In conclusion, this thesis demonstrates that phoneme-level Hierarchical Emotion Distributions offer a powerful and practical solution for fine-grained emotional speech synthesis. It provides both a methodological blueprint and a validation for controllable prosody generation, which sets the stage for future advancements in expressive, interpretable, and human-centric speech technologies.

References

Chen, H., Chen, R., & Hirschberg, J. (2024). EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 8170–8180). Association for Computational Linguistics. Retrieved from `https://doi.org/10.18653/v1/2024.emnlp-main.466` doi: 10.18653/v1/2024.emnlp-main.466

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1–2), 5–32. Retrieved from `https://doi.org/10.1016/S0167-6393(02)00071-7` doi: 10.1016/S0167-6393(02)00071-7

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Retrieved from `https://doi.org/10.48550/arXiv.1810.04805` doi: 10.48550/arXiv.1810.04805

Diatlova, D., & Shutov, V. (2023). *Emospeech: Guiding fastspeech2 towards emotional text to speech.* Retrieved from `https://doi.org/10.48550/arXiv.2307.00024` doi: 10.48550/arXiv.2307.00024

Du, C., & Yu, K. (2021). Rich prosody diversity modelling with phone-level mixture density network. In *Proceedings of interspeech 2021* (pp. 3136–3140). Retrieved from `https://doi.org/10.21437/Interspeech.2021-802` doi: 10.21437/Interspeech.2021-802

Ellinas, N., Christidou, M., Vioni, A., Sung, J. S., Chalamandaris, A., Tsiakoulis, P., & Mastorocostas, P. (2023). Controllable speech synthesis by learning discrete phoneme-level prosodic representations. *Speech Communication*, *146*, 22–31. Retrieved from `https://doi.org/10.1016/j.specom.2022.11.006` doi: 10.1016/j.specom.2022.11.006

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. Retrieved from `https://doi.org/10.1109/TAFFC.2015.2457417` doi: 10.1109/TAFFC.2015.2457417

Gobl, C. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1–2), 189–212. Retrieved from `https://doi.org/10.1016/S0167-6393(02)00082-1` doi: 10.1016/S0167-6393(02)00082-1

Goudbeek, M., Goldman, J.-P., & Scherer, K. R. (2009). Emotion dimensions and formant position. In *Proceedings of interspeech 2009* (pp. 1575–1578). Retrieved from `https://doi.org/10.21437/Interspeech.2009-469` doi: 10.21437/Interspeech.2009-469

Guo, Y., Du, C., Chen, X., & Yu, K. (2023). *Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance.* Retrieved from `https://doi.org/10.48550/arXiv.2211.09496` doi: 10.48550/arXiv.2211.09496

Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., ... Pang, R. (2018). *Hierarchical generative modeling for controllable speech synthesis.* Retrieved from `https://doi.org/10.48550/arXiv.1810.07217` doi: 10.48550/arXiv.1810.07217

Inoue, S., Zhou, K., Wang, S., & Li, H. (2024a). Fine-grained quantitative emotion editing for speech generation. In *Proceedings of the 2024 asia pacific signal and information processing association annual summit and conference (apsipa asc)* (pp. 1–6). Retrieved from `https://doi.org/10.1109/APSIPAASC63619.2025.10848721` doi: 10.1109/APSIPAASC63619.2025.10848721

Inoue, S., Zhou, K., Wang, S., & Li, H. (2024b). Hierarchical emotion prediction and control in text-to-speech synthesis. In *Proceedings of icassp 2024 - 2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 10601–10605). Retrieved from https://doi.org/10.1109/ICASSP48485.2024.10445996 doi: 10.1109/ICASSP48485.2024.10445996

Jiang, Y., Li, T., Yang, F., Xie, L., Meng, M., & Wang, Y. (2024). Towards expressive zero-shot speech synthesis with hierarchical prosody modeling. In *Proceedings of interspeech 2024* (pp. 2300–2304). Retrieved from https://doi.org/10.21437/Interspeech.2024-2506 doi: 10.21437/Interspeech.2024-2506

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. Retrieved from https://doi.org/10.1037/0033-2909.129.5.770 doi: 10.1037/0033-2909.129.5.770

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*(2), 237–265. Retrieved from https://doi.org/10.3758/s13423-019-01701-x doi: 10.3758/s13423-019-01701-x

Kim, J., Kong, J., & Son, J. (2021). *Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech.* Retrieved from https://doi.org/10.48550/arXiv.2106.06103 doi: 10.48550/arXiv.2106.06103

Kiritchenko, S., & Mohammad, S. M. (2017). *Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation.* Retrieved from https://doi.org/10.48550/arXiv.1712.01765 doi: 10.48550/arXiv.1712.01765

Lai, C. (2009). Perceiving surprise on cue words: Prosody and semantics interact on *right* and *really*. In *Proceedings of interspeech 2009* (pp. 1963–1966). Retrieved from https://doi.org/10.21437/Interspeech.2009-475 doi: 10.21437/Interspeech.2009-475

Lei, Y., Yang, S., & Xie, L. (2020). *Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis.* Retrieved from https://doi.org/10.48550/arXiv.2011.08477 doi: 10.48550/arXiv.2011.08477

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of interspeech 2017* (pp. 498–502). Retrieved from https://doi.org/10.21437/Interspeech.2017-1386 doi: 10.21437/Interspeech.2017-1386

Morrison, M., Rencker, L., Jin, Z., Bryan, N. J., Caceres, J.-P., & Pardo, B. (2021). *Context-aware prosody correction for text-based speech editing.* Retrieved from https://doi.org/10.48550/arXiv.2102.08328 doi: 10.48550/arXiv.2102.08328

Oh, H.-S., Lee, S.-H., & Lee, S.-W. (2024). Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2654–2666. Retrieved from https://doi.org/10.1109/TASLP.2024.3395994 doi: 10.1109/TASLP.2024.3395994

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. Retrieved from https://doi.org/10.1016/S0167-6393(02)00084-5 doi: 10.1016/S0167-6393(02)00084-5

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... Saurous, R. A. (2018). *Towards end-to-end prosody transfer for expressive speech synthesis with tacotron.* Retrieved from https://doi.org/10.48550/arXiv.1803.09047 doi: 10.48550/arXiv.1803.09047

Tae, J., Kim, H., & Kim, T. (2022). *Editts: Score-based editing for controllable text-to-speech.* Retrieved from https://doi.org/10.48550/arXiv.2110.02584 doi: 10.48550/arXiv.2110.02584

Tits, N. (2022). Controlling the emotional expressiveness of synthetic speech: A deep learning approach. *4OR*, *20*(1), 165–166. Retrieved from https://doi.org/10.1007/s10288-021-00473-2 doi: 10.1007/s10288-021-00473-2

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). *Wavenet: A generative model for raw audio.* Retrieved from https://doi.org/10.48550/arXiv.1609.03499 doi: 10.48550/arXiv.1609.03499

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). *Attention is all you need.* Retrieved from https://doi.org/10.48550/arXiv.1706.03762 doi: 10.48550/arXiv.1706.03762

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... Saurous, R. A. (2017). *Tacotron: Towards end-to-end speech synthesis.* Retrieved from https://doi.org/10.48550/arXiv.1703.10135 doi: 10.48550/arXiv.1703.10135

Xu, Y. (2019). Prosody, tone and intonation. In W. F. Katz & P. F. Assmann (Eds.), *The routledge handbook of phonetics* (pp. 314–356). New York: Routledge.

Yildirim, S., Lee, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., & Narayanan, S. (2004, November). Study of acoustic correlates associated with emotional speech. In *148th meeting of the acoustical society of america.* San Diego, CA. (Conference presentation)

Yoon, T.-J. (2024). How much does the dynamic f0 curve affect the expression of emotion in utterances? *Applied Sciences*, *14*(23), 10972. Retrieved from https://doi.org/10.3390/app142310972 doi: 10.3390/app142310972

Zhang, M., Zhang, H., Tang, E., Ding, H., & Zhang, Y. (2023). Evaluating the relative perceptual salience of linguistic and emotional prosody in quiet and noisy contexts. *Behavioral Sciences*, *13*(10), 800. Retrieved from https://doi.org/10.3390/bs13100800 doi: 10.3390/bs13100800

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, *137*, 1–18. Retrieved from https://doi.org/10.1016/j.specom.2021.11.006 doi: 10.1016/j.specom.2021.11.006

Zhou, K., Sisman, B., Rana, R., Schuller, B. W., & Li, H. (2023). Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, *14*(1), 31–48. Retrieved from https://doi.org/10.1109/TAFFC.2022.3175578 doi: 10.1109/TAFFC.2022.3175578

Appendices

A Declaration of AI use

I hereby affirm that this Master thesis was composed by myself, and that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced.

During the preparation of this thesis, I used ChatGPT (GPT-4, OpenAI, version June 2024) for the following support purposes:

Grammar refinement and sentence restructuring across various parts of the thesis, without altering the substantive content or argumentation.

Debugging Python code related to the implementation of HED vector injection and emotion control mechanisms described in Chapter 3 (Methodology), including architectural integration and inference script adaptation.

Brainstorming alternative phrasings and rewording to improve argument clarity in the Discussion chapter.

Summarizing prior literature during the early stages of Chapter 2 drafting to guide manual literature organization and citation gathering.

All outputs from the AI tool were critically reviewed, manually verified, and substantially revised by me to ensure academic integrity, domain-specific correctness, and consistency with my own research contributions.

I did not use AI tools to generate hypotheses, determine experimental designs, interpret evaluation results, or draw conclusions. All data analysis, result interpretation, and evaluative judgment reflect my own understanding, reasoning, and domain expertise.

Name: Qiyan Huang

Date: 11/06/2025