university of groningen / campus fryslân

From Text to Feeling
Fine-Tuning FastSpeech2 for Emotion Expression

Fabiènne Nicolaij

university of groningen / campus fryslân

University of Groningen

From Text to Feeling
Fine-Tuning FastSpeech2 for Emotion Expression

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
Dr. V. Verkhodanova (Voice Technology, University of Groningen)
and
Dr. M. Coler (Voice Technology, University of Groningen)

Fabiènne Nicolaij (s4983416)

August 18, 2025

Contents

Acknowledgments

I would like to thank my supervisor Vass Verhodanova for her guidance, support, and most of all her patience throughout this research. I would also like to thank my boyfriend Kristian for proofreading the text and keeping me motivated. Finally, I would like to thank my classmates and the VoiceTech team for the unforgettable time in Leeuwarden and beyond.

Abstract

This study investigates whether FastSpeech 2 can produce emotionally expressive speech without explicit use of emotion embeddings. Two training strategies were compared: (1) training FastSpeech 2 from scratch using an English subset of the Emotional Speech Dataset (ESD), and (2) fine-tuning a FastSpeech 2 model pre-trained on the neutral LJSpeech corpus with the same emotional ESD subset. The emotions available in this dataset are “neutral, “happy”, “angry”, “sad”, and “surprise”. The two models were evaluated using spectral fidelity (Mel-Cepstral Distortion; MCD) and prosodic accuracy (Mean Absolute Error; MAE). Pitch, duration, and intensity were analysed as they are the fundamental prosodic features in FastSpeech 2.

The results demonstrate that the scratch-trained model (setup 1) outperforms the fine-tuned model (setup 2) across all metrics. Setup 1 showed lower MCD and lower MAE across all emotional categories, indicating better reproduction of emotional prosody. Setup 2 struggled to reproduce expressive prosody, particularly for high-intensity emotions, suggesting that pre-training on neutral speech introduces a bias that limits the model’s ability to adapt to emotional variability. In addition, prosodic patterns in setup 1 closely reflected those measured in the ground-truth data, whereas setup 2 often produced flattened pitch contours and shortened durations.

These findings indicate that FastSpeech 2 can implicitly learn and reproduce emotion-specific prosody when trained directly on expressive data. Training from scratch, even on a small dataset, can outperform fine-tuning from a neutral model. This highlights the importance of matching training data to task objectives in emotional speech synthesis and suggests that emotional speech synthesis benefits from training on expressive datasets rather than large neutral ones.

1 Introduction

1.1 Introduction

Speech is not only used to convey words but also to express emotions, intentions, and social information. Emotion plays an important role in how listeners perceive and interpret spoken messages. Developing speech synthesis systems that can express emotion naturally continues to be difficult. This difficulty largely stems from the multiple acoustic dimensions conveyed when an emotion is expressed, such as pitch, intensity, duration, and timbre (Eyben, Wöllmer, & Schuller, 2010; Wu et al., 2022).

Traditional concatenative and parametric TTS systems struggled to capture emotional expression as they relied on fixed speech units and offered limited prosodic control, often resulting in a robotic or monotonous speech (Taylor, 2009). Tacotron 2 (Y. Wang et al., 2017) improved naturalness and prosodic variation through end-to-end learning (Shen et al., 2017). However, these autoregressive models, which generate speech frame by frame, tend to have slower inference and reduced stability for longer speech outputs. They also encounter difficulties modelling all the different ways prosody can change (Shen et al., 2017).

To address these issues non-autoregressive models, which generate all speech frames in parallel, such as FastSpeech (Ren et al., 2019) and and FastSpeech 2 (Ren et al., 2020) have been introduced. In particular, FastSpeech 2 enables fast and reliable synthesis, coupled with direct control of prosodic features, pitch, intensity, and duration, that are closely associated with emotional speech (Ren et al., 2020). For instance, happiness and anger are typically associated with higher pitch and intensity, while sadness has lower pitch and a slower rate (Gobl & Chasaide, 2003; Pell, Paulmann, Dara, Alasseri, & Kotz, 2009). Some prosodic features depend on cultural context, but many emotional cues are shared across languages (Pell et al., 2009). FastSpeech 2's variance adaptor models these prosodic features for fine control. Yet, without emotion labels or conditioning the model struggles to achieve rich emotional expressiveness in synthesis (Ren et al., 2020).

Most existing research on emotional TTS using FastSpeech 2 includes explicit emotional control, often through embeddings derived from labelled data or reference encodes (Lee, Rabiee, & Lee, 2017). Although effective, these methods rely on expensive and sometimes inconsistent emotional annotations. They also raise concerns about whether emotional expression can be generalised effectively to unseen speakers or domains without external emotional cues (S. Zhang, Mehrish, Li, & Poria, 2025). The capability of FastSpeech 2 to learn emotional variation implicitly from unlabelled expressive speech data remains largely unexplored. Some evidence suggest that neural TTS systems can learn affective traits by statistically capturing prosodic patterns (Wu et al., 2022). However, it remains unclear how well emotional prosody is learned without labels aligns with human perception of real emotion.

1.2 Research Question and Hypothesis

As stated, supervised approaches for emotional TTS typically rely on explicit emotion labels and embeddings (Auzawa, Iwasawa, & Matsuo, 2018; Wu et al., 2022; S. Zhang et al., 2025), which are limited by factors such as annotation cost and inconsistency (Cowie & Cornelius, 2003; Eyben et al., 2015). FastSpeech 2 offers an alternative through the variance predictors for pitch, duration, and

intensity, allowing prosodic control without explicit emotion annotations (Ren et al., 2020). These parameters are important cues in the perception of vocal emotion (Banse & Scherer, 1996; Mozziconacci, 2002) and can convey emotion without changes in the lexical content (Cowie et al., 2001; Gobl & Chasaide, 2003).

This thesis investigates to what extent FastSpeech 2 can reproduce emotional prosody without explicit emotion embeddings, compared to the ground-truth. Based on prior findings (C. Cui et al., 2021; Diatlova & Shutov, 2023) it is expected that the variance predictors alone can internalise affect-specific prosodic patterns, leading to outputs with emotion specific pitch, duration, and intensity. Studies demonstrate that emotional content can be inferred reliably by listeners from prosody alone (Banse & Scherer, 1996; Cowie et al., 2001; Pell et al., 2009). To test this the ground-truth emotional speech will be compared to a model trained from scratch on expressive data, and a pre-trained model fine-tuned on expressive data.

1.3 Relevance

This research contributes to the field of emotional speech synthesis by providing an empirical evaluation of FastSpeech 2's ability to convey emotional variation using only its internal prosodic predictors, without explicit emotion embeddings. By comparing both scratch-trained and fine-tuned models directly to the ground-truth recordings, offering insights into the capabilities and boundaries of implicit emotion modelling in non-autoregressive architectures. The findings could guide efficient and scalable emotional TTS systems in low-resource contexts where annotated emotional data is scarce. This work also contributes to the ongoing discussion on balancing explicit and implicit approaches to emotion modelling, which could potentially help shape future TTS designs toward models that reduce dependence on labelled emotion embeddings while supporting nuanced and expressive control over prosody (Y. Cui, Wang, Zhao, Zhou, & Chen, 2023; He et al., 2022).

1.4 Thesis Outline

This thesis is structured as follows: Chapter 2 provides an overview of related work in emotional TTS and outlines the main theoretical concepts strengthen the current study. Chapter 3 details the experimental setup, including the dataset, the data preprocessing pipeline, the model configurations, and the evaluation metrics used to evaluate performance. Chapter 4 presents and discusses the results, analysing how effectively FastSpeech 2 captures emotional expressiveness without the use of explicit embeddings. Finally, Chapter 5 concludes the thesis with a summary of the key points and outlining directions for future research in emotional speech synthesis.

2 Related works

2.1 Emotional Speech and Prosody

2.1.1 Emotion in Speech and Prosodic Variation

Emotion in speech is predominantly expressed through prosodic variation. Prosody, which refers to features like intonation, rhythm, stress, and timing, helps listeners understand grammatical structure and emphasis in speech. They also carry social and emotional meaning (Ladd, 2008; Xu, 2019). Emotional prosody is typically analysed by changes in pitch (the fundamental frequency, which is a physical correlate of perceived intonation), speech rate and timing, loudness (or intensity), and voice quality (e.g. tenseness, breathiness, or creakiness) (Campbell & Mokhtari, 2003; Gobl & Chasaide, 2003). These features can give insight into a speaker's emotional state, even when the words used are not necessarily emotional or familiar (Cowie et al., 2001; Larrouy-Maestri, Poeppel, & Pell, 2024). A small change in timing or intonation can change the perception from sympathy, to confidence, sarcasm, or hesitation, shaping both the social and emotional tone of a conversation (Goudbeek & Scherer, 2010; Scherer, 2003). Emotional prosody not only makes speech more interesting for the listener to hear, it conveys communicative intent, shaped by internal states, context, and speaker-listener relationships (Cowie et al., 2001; Scherer, 2009). Voice quality adds another expressive layer, as a tense voice could imply anger or urgency, while a soft, breathy tone may suggest sadness or warmth (Campbell & Mokhtari, 2003; Gobl & Chasaide, 2003).

Different emotional states tend to differ in their prosodic patterns, making it possible to distinguish them in speech. Anger is often found to have elevated pitch, wide pitch range, higher intensity, and faster speech rate. Sadness, in contrast, has a lower pitch, lower intensity, slower tempo, and longer pauses (Banse & Scherer, 1996; Scherer, 2003). Happiness tends to be expressed through a wider pitch range and faster rhythm, while fear could be a combination of high pitch with erratic pacing and sharp articulation (Juslin & Laukka, 2003). Even though these patterns are not fully universal, they can be seen across different languages and cultures. Emotional tone can be rightfully detected across unfamiliar languages, demonstrating that prosodic emotion cues have cross-linguistic stability (Laukka & Elfenbein, 2020; Pell et al., 2009). These findings indicate that prosodic variation in emotional speech is shaped by both human biology and universal features of interaction (Scherer, 2003). Nevertheless, variations are observed even within the same emotion, often attributed to cultural norms, gender roles, and language-specific phonological structures which influence how emotions are expressed and perceived. In tonal languages, for instance, pitch is partly constrained by lexical tone, resulting in a greater use of intensity or voice quality to convey emotion (Chang et al., 2023). Gender, in turn, appears to influence both production and perception of prosody. Female listeners are more sensitive to subtle emotional shifts, while male listeners may depend on more pronounced acoustic cues (X. Wang, Fang, & Ding, 2024). These sociolinguistic and contextual factors reinforce the idea that emotional prosody is a dynamic and complex communicative tool, they do not always follow fixed prosodic patterns (Larrouy-Maestri et al., 2024; van Rijn & Larrouy-Maestri, 2023). To better understand how prosody conveys emotional meaning, it is useful to examine how emotional itself is conceptualised in different models.

2.1.2 Models of Emotion

Several models have been proposed to describe and classify emotion, this section covers how these models relate to speech prosody. Emotion is commonly described using either categorical or dimen-

sional approaches. Categorical models treat emotions, such as anger, happiness, fear, and sadness, as discrete states. Each state is associated with relatively stable and recognisable prosodic patterns (Ekman, 1992; Izard, 1993). Models such as these are easily applicable in annotation and recognition tasks, making them widely used in experimental and computational research. Dimensional models, in contrast, describe emotions as positions within a continuous space. The most frequently used dimensions for this space are arousal, ranging from calm to excited, and valence, from negative to positive (Goudbeek & Scherer, 2010; Russell, 1980; Scherer, 2003). This framework is useful for capturing subtle emotional variation, such as when a speaker shifts from mild irritation to full anger. Prosodic features like speech rate, pitch, and intensity are more closely linked to arousal than to valence, which reflect how energy levels are expressed through speech (Scherer, 2003). Emotions such as anger, happiness, fear, and surprise are typically classified as high-intensity due to their association with elevated arousal and vocal energy, while emotions like sadness or boredom are associated with low-intensity, marked by subdued acoustic features (Juslin & Laukka, 2003; Scherer, 2003).

Hybrid models combine elements of both categorical and dimensional frameworks by mapping categorical emotions to specific locations within a dimensional space. This allows for more flexible annotations and better representation of nuanced expressions (Cowie et al., 2001; Juslin & Laukka, 2003). These models are effective for analysing how emotions vary across the duration of speech. The Component Process Model expands on this by framing emotion as a sequence of appraisals that drives both vocal and physiological responses (Scherer, 1986, 2009). This model helps explain both intra- and interspeaker variation in how emotions show through prosody. Each model provides unique insights and, taken together, offer a better understanding of how emotional signals are produced and perceived.

In the present study a categorical approach is adopted, focussing on discrete emotion labels to guide and evaluate prosodic modelling in FastSpeech 2. This allows direct mapping of emotion-specific prosody onto model performance, connecting theory with practical implementation in emotionally expressive TTS.

2.1.3 Challenges in Emotional Prosody Interpretation

Interpreting emotional prosody is a challenge both perceptual and linguistic, even with established models. Emotion perception is often ambiguous, as it can be interpreted differently from one individual to the next, even when the acoustic cues are prototypical of a specific emotion. This difference can result from cultural background, individual listening habits, or expectations (Laukka & Elfenbein, 2020; Pell et al., 2009). Tools like FEELTRACE, which track emotional perception in real time, demonstrate that listeners shift their impressions of valence and arousal over the course of an utterance, making single-label annotations an oversimplification of listeners' actual perceptual experience (Cowie et al., 2001). At a linguistic level, prosodic signals used for grammar or discourse, such as a question intonation or emphasis, often overlap with emotional cues. This can diminish the line between structure and affect (Hirst & di Cristo, 1998; Ladd, 2008). The role of voice quality is similarly ambiguous. Features such as breathiness or creakiness might signal emotion. They can, however, also be tied to speaker style or linguistic contrast (Campbell & Mokhtari, 2003; Gobl & Chasaide, 2003).

The challenges mentioned can be emphasised when prosodic features do not align. For example, a speaker might use high pitch, suggesting excitement, alongside a slow tempo, often tied to sadness, leading to ambiguous emotional impressions (Goudbeek & Scherer, 2010). These inconsistencies

show the limitations of analysing prosody without considering broader contextual factors. Mirroring this challenge, differences in methodology make it difficult to compare findings across studies. The use of acted or spontaneous speech, as well as variation in the annotation and presentation of emotional stimuli, can lead to inconsistent results and restrict the extent to which the findings can be generalised (Larrouy-Maestri et al., 2024). Commonly used frameworks are still being refined, however, the lack of uniform emotion categories, the need for clearer labelling standards, and improvement in how emotional changes are captured over time remain important areas for ongoing research and development (Larrouy-Maestri et al., 2024; Themistocleous, 2025).

2.2 Emotional Speech Synthesis

2.2.1 Introduction to Emotional Speech Synthesis

Emotional Speech Synthesis (ESS) aims to bridge the gap between natural and synthetic speech by generating speech that not only conveys linguistic content but also mirrors the emotional tone typical of human communication. Synthetic voices are increasingly used in daily technologies. Applications span a wide range of domains, including audiobooks, video games and costumer-facing services. Many of these voices rely on flat, neutral prosody, creating a growing demand for systems that offer expressive and emotionally adaptive interactions (Barakat, Turk, & Demiroglu, 2024; Eyben et al., 2010). As emotional appropriateness can enhance user engagement and satisfaction (Schröder, 2004; Skerry-Ryan et al., 2018).

Early Text-To-Speech (TTS) systems focused on producing clear and intelligible speech. This often resulted in monotonous or unnatural prosody. The systems created produced speech that was perceived as robotic, reducing their usefulness in interactions where emotional expression helps shape intent or meaning (Taylor, 2009). However, the rise of data-driven models, especially those using deep neural networks, have increased the ability of synthetic voices to produce expressive and more natural-sounding speech (Stanton et al., 2018). They allow for finer-grained control of prosodic parameters such as pitch, duration, and intensity, which are essential to conveying affective intent. Emotional TTS aims to enhance the expressiveness not by simply adding surface-level variation, but by generating speech that reflects deeper affective states aligned with the communicative context. Emotional cues, such as urgency, calmness, sympathy, or enthusiasm, help convey speaker intent and shape how the message is perceived. This enables synthetic voices to engage listeners in a way that feels more natural and socially attuned (Cowie et al., 2001; Schröder, 2004). In domains like education or mental health, such prosodic subtlety can significantly influence how information is received and interpreted.

2.2.2 Advances and Challenges

ESS has progressed significantly in recent years, the adoption of neural networks architectures in TTS allowed for precise modulation of prosodic cues. Models such as Tacotron 2 and FastSpeech 2 offer mechanisms for controlling pitch, duration, and intensity, parameters which are closely linked to emotional expression (Ren et al., 2020; Skerry-Ryan et al., 2018). Style tokens, learned vectors that represent different speaking styles or patterns in speech, and latent emotion embeddings build upon these advances by facilitating the modelling of expressive variation without the use of labelled emotion data (Stanton et al., 2018). This reduced reliance on emotion annotations, which can be costly and inconsistent, making expressive synthesis adaptable across domains. Approaches like ED-TTS and EmoMix model emotion at multiple temporal scales and enable dynamic blending of emotional intensity (Tang, Zhang, Cheng, Xiao, & Wang, 2024; Tang, Zhang, Wang, Cheng, & Xiao, 2023a).

Together these methods form a comprehensive framework for emotional stylisation, bringing emotional TTS systems closer to enabling modelling of emotional subtleties that reflect the complexities of real human speech.

These advances are, however, not without challenges. Limitations in data availability, particularly the lack of high-quality datasets that cover a range of emotions, speakers, and linguistic contexts, have been shown to reduce generalisability and robustness of the model (Barakat et al., 2024). Many available datasets rely on acted speech or simplified emotion labels, typically reduced to basic categories like "happy", "sad", and "neutral", which often fails to capture authentic prosodic variation (Cowie et al., 2001; Larrouy-Maestri et al., 2024). Natural emotion expression is dynamic and context dependent, emotions blend, overlap, and change over time. Categorising emotions in such conditions into a discrete form leads to mislabelling. This challenge is further complicated by perceptual ambiguity. Listeners often struggle to distinguish between emotions with similar acoustic features, confusing, for example, surprise with excitement (Barakat et al., 2024; van Rijn & Larrouy-Maestri, 2023). Yet another challenge is the design of reliable evaluation frameworks. As stated, emotional perception can vary considerably across individuals and cultures. Standard metrics like MOS (Mean Opinion Score) or simple emotion classification often fail to reflect the perceived complexity of emotions (Larrouy-Maestri et al., 2024). Tools like FEELTRACE offer alternatives by tracking continuous valence and arousal during listening, but aligning synthesised speech onto these perceptual data remains difficult as synthesised speech cannot make the switch from on emotion to the next as clearly or naturally (Cowie et al., 2001).

Attempts have been made to circumvent these limitations, with varying degrees of success. Cross-lingual and zero-shot emotion transfer methods, including systems like METTS, try to do so by synthesising emotion in languages or emotions unseen during training. They are, however, constrained by a need for large datasets (Li et al., 2023; Zhu et al., 2024). Another possible solution is context-aware synthesis, where prosody adapts based on sentence structure, discourse, function or dialogue history (Liu, Yifan, Ren, Yin, & Li, 2024). To improve the clarity of emotional intent and lessen the ambiguity, visual cues (facial expressions or gestures) can be added to the emotional audio. This multimodal synthesis gives the listener additional contextual information, increasing their ability to accurately interpret the intended emotion (Galdino, Matos, Svartman, & Aluisio, 2025). While these advances continue to narrow the gap between synthesised and human emotional expression, emotion itself remains a complex phenomenon that may never be fully captured by computer models.

2.3 Emotion Modelling in Neural TTS

2.3.1 Emotion Representation and Annotation

Emotion modelling in neural TTS systems involves identifying and shaping emotional cues to make synthesised speech more engaging and human-like. A common method is with the use of categorical emotion embeddings, where each label is linked to a learnable vector that informs the synthesis process. In MeTTS, categorical embeddings are embedded and concatenate with decoder inputs, allowing the model to synthesise speech that reflects the intended emotional tone (Zhu et al., 2024). Similar techniques are used in expressive variants of FastSpeech, where embeddings guide the generation of pitch, duration, and intensity by conditioning the model at the variance adaptor or decoder (Tang et al., 2023a). While these methods are quite efficient and straightforward to implement, they require high-quality and well-balanced emotional datasets for optimal performance. If the annota-

tions are sparse or unreliable the model risks overfitting to superficial prosodic features that fail to generalise or express subtle affective distinctions (Barakat et al., 2024; van Rijn & Larrouy-Maestri, 2023).

Categorical emotion modelling has limitations in representing the full range of human emotion. It assumes that emotions are discrete and mutually exclusive, which contradicts psychological models of emotion that suggest that affective states are dynamic and often overlap in nature (Ekman, 1992; Russell, 1980). Emotions such as curiosity, boredom, or ambivalence frequently pop-up in speech, however, they rarely fit neatly into fixed classes. And once more, emotional annotation is frequently affected by inconsistencies due to subjective perception and cultural differences in expressing emotion (Cowie et al., 2001; Laukka & Elfenbein, 2020).

To address these limitations, researchers have begun modelling emotion as continuous rather than discrete, using dimensional models that situate emotions along axes such as arousal and valence. Taking inspiration from frameworks like the Circumplex Model of affect (Russell, 1980) and the component Process Model (Scherer, 2009), emotions are not seen as fixed categories, but as fluid states that vary in intensity and are shaped by cognitive and contextual variables. Continuous values for arousal and valence can be extracted from speech datasets using acoustic inference models or human ratings (Cowie et al., 2001; Larrouy-Maestri et al., 2024), which then guide neural generation through conditioning layers (Tang et al., 2024; Zhu et al., 2024). This approach allows for finer emotional control and the generation of intermediate affective states, such as mild annoyance or intense joy. This level of precision is particularly useful in applications such as storytelling and dialogue systems, where emotional tone evolves with context. However, continuous models often struggle with generalisability due to their reliance on subjective annotations and are less interpretable across cultures and languages, as emotional categories and intensities may not map uniformly (Larrouy-Maestri et al., 2024; van Rijn & Larrouy-Maestri, 2023). Nonetheless, these dimensional models offer a scalable and psychologically grounded framework for capturing and replicating the richness and variability of emotional expression in synthetic speech.

2.3.2 Embedding and Controlling Emotion in TTS Models

Controlling emotional expression in neural TTS models depends on how emotion is encoded and introduced into the model's architecture. A commonly used method is to add emotional information as auxiliary input, this allows the model to control prosodic features according to a given emotional label. In category-based systems each emotion label (e.g. "happy", "angry", "sad") is assigned to a learnable embedding vector (Barakat et al., 2024). This vector is then integrated into the model pipeline, typically through concatenation or decoder inputs. Emotion-conditioned Tacotron, for example, integrates the emotion embeddings directly into the decoder layer (Lee et al., 2017), whereas systems with an FastSpeech-based architecture guide the embeddings through the variance adaptor to influence the prosodic predictors (Ren et al., 2020). When using style tokens, another strategy is employed. They open up an unsupervised pathway as latent embeddings and are learned without reliance on labelled data (Stanton et al., 2018). These tokens, and their successors such as Global Style Tokens (GSTs), allow models to generalise expressive variation and support zero-shot transfer across speakers (J. Zhang, Wushouer, Tuerhong, & Wang, 2023). Recent advances use vector quantisation techniques, which allows models to discretely encode latent emotional clusters into discrete units that can be selectively applied during synthesis (Tang et al., 2024). These embeddings offer intuitive and easy control. However, the effectiveness of each approach depends on the specific use case and

available data. Category-based methods suit explicit emotion control, while style tokens and vector quantisation offer greater flexibility, but they all may struggle to capture subtle emotional nuances if the training data is not sufficiently diverse or balanced.

Advanced architectures attempt to overcome these challenges by modelling emotion independently from speaker and linguistic features, often using multi-scale or hierarchical approaches. ED-TTS uses hierarchical embeddings to model emotion at both local (phoneme or word) and global (utterance) level, enabling more precise control over emotional intensity and timing (Tang et al., 2024). Similarly, models like EmoDiff (Guo, Du, Chen, & Yu, 2022) and EmoMix (Tang et al., 2023a) use diffusion-based and prototype-driven mechanisms, respectively, to generate expressive speech or mixed emotion embeddings. Diffusion-based mechanisms gradually transform random noise into realistic and expressive speech by diffusing emotion information over time (Guo et al., 2022). Prototype-driven mechanisms combine known "prototype" or basic emotion embeddings and tries to blend them to create nuanced emotions (Tang et al., 2024). These techniques increase prosodic diversity while maintaining control over emotional expression. Some systems use reference encoders that learn to extract emotional features from sample utterances, supporting low-resource transfer of expressive style across various speakers and languages (Zhu et al., 2024). Other models separate speaker-specific and prosodic features, such as speaker identity, pitch, or rhythm, from emotional embeddings to reduce confounding influences and improve generalisation (X. Wang et al., 2024). Although these methods offer expressive flexibility, they also raise issues of interpretability and stability. Emotion embeddings often capture overlapping features, making it difficult to get fine-grained control without introducing unwanted artefacts. Nonetheless, embedding-based methods have become a foundation in neural TTS pipelines due to their scalability, efficiency, and compatibility with modern architectures.

2.4 FastSpeech 2 and Emotion Modelling

2.4.1 Architecture and Prosodic Control in FastSpeech 2

FastSpeech 2 was not originally designed for emotional synthesis, yet its fast inference, prosody control mechanisms, and flexible architecture have made it a widely used foundation for emotion-aware extensions. Unless otherwise noted, the following description of FastSpeech 2 is mainly drawn from (Ren et al., 2020), which provides a comprehensive account of its architecture and prosody modelling techniques.

Emotional nuances are not inherently modelled into synthesised speech, though the system provides a stable backbone for integrating more advanced emotional control strategies. As its predecessor FastSpeech (Ren et al., 2019), it uses feed-forward Transformer (FFT) blocks and a non-autoregressive framework, enabling parallel processing of speech frames for faster synthesis and better stability.

FastSpeech 2 consists of three core components, an encoder, a variance adaptor, and a decoder, each serving a distinct role in the synthesis pipeline. The encoder turns phoneme sequences into context-aware embeddings, using layers of self-attention and convolution. The decoder maps these time-aligned embeddings to mel-spectrogram frames, which will serve as the intermediate acoustic representation used to synthesise audio signals. What sets FastSpeech 2 apart is what happens during this process, the variance adaptor gives the model direct control over how prosody is handled. During training it is conditioned on ground-truth pitch, intensity, and duration values to learn how these features vary naturally in expressive speech. Then during inference FastSpeech 2 uses the learned

predictors to estimate these prosodic features based on the input text. These predicted values are embedded through learned embeddings and integrated into the hidden representation before being passed to the decoder, allowing the model to generate speech with specific prosodic characteristics.

The variance predictors, for pitch, intensity, and duration, share a common architecture consisting of two convolutional layers to process the sequence, followed by layer normalisation, dropout, and ending in a linear output layer. They are trained using mean squared error (MSE) loss to optimise pitch and intensity predictions and mean absolute error (MAE) loss to optimise duration predictions. Separately, the final mel-spectrogram is trained using the same losses to minimise the difference between predicted and target spectrogram frames. Duration prediction is done using a logarithmic scale to reduce variance and enhance training stability. This configuration is particularly effective to directly address the one-to-many mapping issue common in speech synthesis. Instead of relying on a teacher model, like in the original FastSpeech model (Ren et al., 2019), FastSpeech 2 uses the prosodic features learned to encode natural variation into its predictions. As a result, the training process becomes more straightforward and leads to more natural and expressive speech.

2.4.2 Emotion Modelling Challenges and Adaptations

Many studies have used FastSpeech 2 as a foundation to generate more emotionally expressive speech. This is typically done by adding emotion-aware modules onto the base model. EmoSpeech (Diatlova & Shutov, 2023) does this by placing emotion conditioning blocks before and after the variance adaptor. These added transformer layers help the model learn how to shape emotional intensity across phonemes, enhancing both expressivity and controllability of the speech output. Similarly, the Japanese model by Ikeda and Markov 2024 has added attention mechanisms, which are guided by emotion labels, surrounding the variance adaptor. Compared to the FastSpeech 2 base model this modification led to improvements in both speech naturalness and emotional expressiveness, as they allow emotion to be integrated at points in the pipeline that directly affect pitch, duration, and intensity.

Other methods focus on the specific emotion categories and how intense those emotions are expressed. Emovie (C. Cui et al., 2021) uses embedded emotion labels to guide the speech generation process. The use of explicit labels helps the model to generalise within each category for consistent output, however, the intensity cannot be easily varied. Emoq-TTS (Im, Lee, Kim, & Lee, 2022) builds on this by splitting emotion intensity into levels and assigns each phoneme a score to offer more fine-grained control. The emotional intensity within an utterance can be varied and certain words or syllables can be emphasised more strongly to convey nuanced emotions. Similarly, CASEIN (Y. Cui et al., 2023) uses both explicit and implicit controls to shape emotional intensity more smoothly. This hybrid approach supports stable category control while allowing subtle expressive shifts. This makes it possible to generate blended emotions not present in the training data.

A growing body of work is looking at how modelling structure and context interact with emotion. Cornille et al. 2022 created a system with sliders to adjust prosodic features like pitch, energy, and rate. This is not only useful but also shows that the model can adapt to user input. On the other hand, Inoue et al 2024 introduced predictors that take into account information at the phoneme, word, and sentence levels. Using these linguistic levels allows prosody to match the structure and context of the utterance more closely. The effect is more coherent and natural emotional expression.

Some research has turned to more unsupervised representations of emotion. Models such as FleuntTTS (Kim, Um, Yoon, & Kang, 2022), Qi-TTS (Tang, Zhang, Wang, Cheng, & Xiao, 2023b), Emosphere-TS (Cho, Oh, Kim, Lee, & Lee, 2024), and RSET (Shi et al., 2024) use latent features to represent expressive variation, while also introducing architectural adaptations to the FastSpeech 2 backbone to better support emotion modelling. These features are often learned through clustering similar examples, based on their characteristics, or through variational techniques, which map these characteristics into a smooth continuous space that captures underlying patterns of variation, and can capture subtle expressive differences that are not easily defined by emotion categories. In contrast to the explicitly guided models mentioned earlier, these systems rely on emotion representations learned implicitly during training, offering greater flexibility and expressivity.

All of these approaches reflect a drive for more controlled and targeted emotional output. Such control can be achieved through conditioning modules, hierarchical predictors, or learned latent structures. While FastSpeech 2 provides a strong backbone, these additions help guide its prosodic output to better align with human emotional cues. However, hardly any research tests FastSpeech 2 on emotional data without added modifications. If its prosody predictors, pitch, duration, and intensity, can capture emotional cues directly from the data training could be simpler and more flexible. This could be especially valuable for datasets without emotion annotations or for under-resourced languages where such labelling is difficult.

3 Experimental Design

This chapter outlines how the experiment was set up to explore how effectively emotional prosody can be reproduced through text-to-speech (TTS) systems by comparing two training strategies: training from scratch and fine-tuning a pre-trained model. It is structured around three components: data preprocessing, model configuration and training, and performance evaluation. Preprocessing involved adapting an existing LibriTTS pipeline to accommodate the structure of the ESD dataset. Model training included architectural modifications, training parameter adjustments, and regularisation techniques to address the challenges posed by the limited size of emotional speech data. Both models were trained using consistent configurations to ensure comparability.

3.1 Model Implementation

This study employed an openly available implementation of the FastSpeech 2 model to generate emotional speech[1]. Although this implementation differs slightly from the original FastSpeech2 (Ren et al., 2020), its architecture and design remain largely similar. Notable differences include the use of phoneme-level pitch and energy predictions, rather than the original frame-level, allowing more accurate mapping of prosodic features to linguistic units (Ren et al., 2020). Pitch and energy features were normalised, improving the model's performance by enhancing training stability and enabling more consistent and natural synthesis. Additionally, linear-scaled bins were applied, instead of log-scale quantisation, potentially improving pitch controllability and synthesis stability.

During training, gradient clipping was employed to stabilise the process by preventing excessively large gradients. Another modification is the use of a 6-layer decoder instead of the original 4-layer one, particularly beneficial for multi-speaker datasets, or in this case multi-emotion dataset. To enhance the quality of the generated Mel spectrograms, a Tacotron-2-styled Post-Net was applied after the decoder, resulting in clearer and more natural speech synthesis. Lastly, the implementation supports both the MelGAN (Kumar et al., 2019) and HiFi-GAN (Kong, Jaehyeon, & Bae, 2020) vocoders. In this setup the HiFi-GAN vocoder was used for synthesis, it produces more natural and realistic speech and has faster inference (Kong et al., 2020). Supported datasets for this implementation are the widely used LJSpeech, LibriTTS, and AISHELL-3, making it versatile for multilingual and expressive speech synthesis tasks.

This implementation was chosen for its accessibility and has been successfully applied in prior work, demonstrating strong performance in expressive and emotional TTS (Kögel, Nguyen, & Cardinaux, 2023; Lenglet, Perrotin, & Bailly, 2023; Udagawa, Saito, & Saruwatari, 2022; Xue et al., 2022).

3.2 Data Preparation and Preprocessing

An emotion-specific dataset was selected consisting of five emotion categories (neutral, happy, angry, sad, and surprise) from a single female speaker, namely speaker 0015. The dataset used is an English subset of the Emotional Speech Dataset (ESD) (Zhou, Sisman, Liu, & Li, 2021). This dataset consist of a balanced representation of multiple emotional states, consistent speaker identity, and high audio quality. The selected set of emotions reflects a range of both high and low arousal states, which can be analysed to show how the model handles diverse prosodic characteristics.

[1] https://github.com/ming024/FastSpeech2

The choice to use an English dataset was motivated by the availability of robust text-to-phoneme conversion tools and pre-trained models for English, as well as the need to minimise cross-linguistic variability in prosody and phoneme alignment. This allowed a more controlled analysis of emotional variation. The ESD provides consistent audio quality and is evenly annotated for emotion. No specific acoustic or demographic features distinguishes the speakers from one another in the English subset of the dataset, making the choice for speaker "0015" effectively random but sufficient for the present study.

The "0015" subset contains 350 utterances per emotional state, making it a total of 1750 utterances or approximately 1.18 hours of speech, on average 16 minutes of spoken material for each emotional state. There is a total word count of 2203 words, of which 997 are unique. Each sample was labelled with its corresponding emotion and pre-processed to meet the input requirements of the FastSpeech 2 model. Processing includes generating phoneme alignments, normalising audio features, and converting transcripts to phonetic representations required by the model pipeline.

The ESD dataset is organised as a main directory containing subdirectories for each speaker. Within each speaker's subdirectory, there is an annotation file that provides the transcripts in a .txt file, as well as additional subdirectories categorised by emotion. Each emotion-specific folder contained the corresponding .wav files for that emotion. Since only one speaker is considered, the designated directory is organized separately. The .txt file was converted to an .csv file separated by tabs and with the corresponding emotion category in the second column. The emotion specific folders only consist of the audio files, lacking sentence transcriptions. Montreal Forced Aligner (MFA)[2] was used to obtain alignments between the utterances and the phoneme sequences, creating .TextGrid files and the transcriptions needed. The .TextGrid files were filed in a directory and separated in emotion-specific folders, placed in the preprocessed_data section of the implementation.

To preprocess the emotional dataset the data processing pipeline originally used for the LibriTTS dataset was adapted. The LibriTTS pipeline was designed to handle a multi-speaker dataset, where each speaker's data are organised into multiple chapters subdirectories. Each subdirectory contains paired audio and transcript files. In contrast, the subset of ESD used provides utterances from a single speaker, speaker 0015, organised into emotion-specific folders, without further chapter segmentation. Because of this structural difference, the chapter iteration in the original pipeline was removed, as it was not relevant. In the modified preprocessing script iterates directly through each emotion-labelled folder, filtering for .wav files and pairing them with corresponding .txt transcript files based on their shared base filename. These pairs are then processed to produce the normalised audio and cleaned phonetic transcripts needed for alignment and model training.

3.3 Model Fine-Tuning Procedure

Two training setups were explored: (1) training FastSpeech 2 from scratch using only the selected ESD subset, and (2) fine-tuning a FastSpeech 2 model pre-trained on the LJSpeech dataset with the same ESD subset. Setup 1 serves as the baseline, providing a reference point to evaluate the benefits of pretraining in setup 2. This ensures that any observed improvements reflect the influence of prior neutral speech.

[2]https://montreal-forced-aligner.readthedocs.io/en/latest/

The process was performed using a modified training script in which the utterances were labelled according to their filename and emotion category, and transcriptions were converted in ARPAbet phonemes using the CMU Pronouncing Dictionary, a phonetic transcription system developed by Advanced Research Projects Agency (ARPA) (Weide, 1998). This phonemes help map prosodic features to linguistic units more effectively during training. For example, the word 'cat' would be represented in ARPAbet as K AE T, explicitly encoding the individual phonetic components of the word.

The limited amount of emotional speech data used increases the risk of overfitting, to reduce this risk several adjustments were made based on preliminary testing. The same adjustments were applied to both the model trained from scratch and the fine-tuned model. The encoder and decoder dropout rates were set to 0.3 and weight decay was set to 0.0001 to encourage more robust generalisation and reduce dependency on specific features or patterns in the training data. The learning rate was reduced by 25% to slow down parameters updates, allowing the model to better process its performance during fine-tuning. Additionally, the first four layers of the encoder were frozen to preserve pre-learned lower-level representations.

The optimal number of training steps was not predetermined. Multiple checkpoints were evaluated based on Mel-Cepstral Distortion (MCD) and the best preforming step was selected for synthesis. Detailed comparisons are presented in the next chapter.

Another adjustment had to be made for setup 2, enabling to skip the optimiser during fine-tuning. The pre-trained model had been trained on neutral speech from a single speaker, the LJSpeech dataset. In contrast, the ESD dataset chosen has a single speaker expressing five distinct emotions, causing a mismatch in data structure. The training implementation groups data by speaker or by emotion depending on the configuration. The skip optimiser option allowed the model to disregard speaker-specific optimiser parameters, ensuring training stability and compatibility while still benefitting from the pre-trained model's learned weights.

These modifications strengthened the model's ability to fine-tune effectively on a small, expressive dataset, enabling better generalisation to different emotional categories. All other hyperparameters and model components followed the defaults provided in the FastSpeech 2 implementation used, unless stated otherwise. The model as described can be found on GitHub[3].

3.4 Synthesising

After fine-tuning,, the model was used to synthesise each of the 1750 utterances originally present in the dataset, 350 per emotional category, for both training from scratch and fine-tuning from LJSpeech. Synthesising the original sentences helped assess how well the model reproduced the prosodic nuances of each emotion. Synthesis was performed using the FastSpeech 2 decoder and HiFi-GAN vocoder.

3.5 Evaluation Metrics and Analysis Methods

Although subjective listening tests are the gold standard for evaluating emotional expressiveness in speech synthesis, this study employed objective acoustic measurements due to practical constraints.

[3]https://github.com/FenNlay/FastSpeech2

Specifically, the limited number of participants available would have led to statistically weak and potentially unrepresentative results. Objective metrics do not directly measure perceived emotion or naturalness, however, they offer interpretable, reproducible, and quantifiable indicators of how accurately the model reproduces the underlying prosodic characteristics associated with different emotional states.

To evaluate the performance of the fine-tuned models in reproducing emotional prosody, a set of objective acoustic metrics was used. Specifically, Mel-Cepstral Distortion (MCD) was used to measure the spectral distance between synthesised and ground-truth speech. This metric offers insight into how closely the generated speech resembles the ground-truth audio. MCD calculates the distance between the Mel-frequency cepstral coefficients (MFCCs) of the original and generated audio samples. A lower MCD score indicates a greater similarity between the two types of audio, reflecting higher synthesise quality (Kubichek, 1993; Toda, Black, & Tokuda, 2007). MCD was calculated for each emotion individually, as well as an overall average score across all categories, for both setups. It is defined as:

$$MCD(X,Y) = \frac{10\sqrt{2}}{\ln 10} \sum_{t}^{T} \sqrt{\sum_{d}^{D} \left(x_t\left(d\right) - y_t\left(d\right)\right)^2}$$

where
X: output mel-cepstrum
Y: target mel cepstrum
T: mel-cepstrum length
D: mel-cepstrum dimension
$x_t\left(d\right)$: dth output mel-cepstrum coefficient in tth frame
$y_t\left(d\right)$: dth target mel-cepstrum coefficient in tth frame

In addition, Mean Absolute Error (MAE) was used to assess the difference between synthesised and ground-truth audio for duration, pitch, and intensity. MAE is less sensitive to large outliers and provides an intuitive measure of difference in the context of speech timing and prosody (Willmott & Matsuura, 2005). Duration analysis measured the total utterance length. Although rhythm encompasses more than duration, this metric provides a baseline approximation of temporal alignment. Pitch analysis focussed on the average absolute deviation of fundamental frequency contours. This feature reflects emotional expressiveness, such as increased pitch in happy speech or lowered pitch in sadness (Gobl & Chasaide, 2003). Intensity was used to examine the loudness variations, reflecting how well the model reproduces the energy dynamics of the target speech. These features were chosen because they correspond to key dimensions of emotional prosody: pitch conveys arousal and affective intent, intensity reflects vocal energy, and duration influences speech rhythm and emphasis (Eyben et al., 2010; Mozziconacci, 2002; Schuller, Batliner, Steidl, & Seppi, 2011).

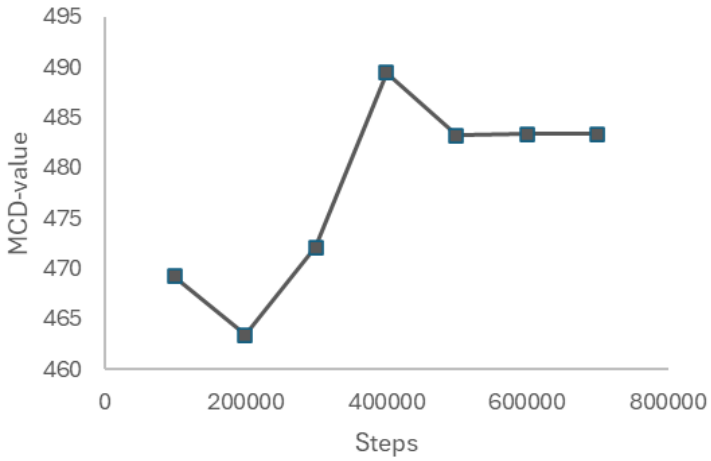

Figure 1: **MCD-values Across Training Steps**. MCD values measured every 100,000 training steps for Setup 1, showing spectral distortion trends and early convergence.

4 Results & Discussion

4.1 Model Convergence

The reliability and generalisability of the model are determined by its performance, as reflected in training and validation loss. As setup 1 forms the base for setup 2, only its training and validation losses are evaluated.

Training loss decreases as the model learns to better predict the target outputs from the training data (Appendix A). The model achieved a mean total training loss of 0.69 after a steep drop early in training, stabilising after around 50,000 steps, where it remained relatively consistent for the remainder of the training. Each of the sub-loss components showed a similar pattern.. Both the mel loss and the mel postnet loss, which measure how accurately the model reconstructs the spectral envelope, averaged at 0.32. This suggests the model was able to effectively learn the spectral features of the training data and reproduce the speech signal of the target audio, matching what is observed in the original FastSpeech 2 implementation (Ren et al., 2020). Additionally, the model learned to predict the prosodic features with reasonable precision as indicated by the low duration, pitch, and energy losses, 0.01, 0.02, and 0.02 respectively. Taken together, these results suggest a reliable foundation for generating expressive, natural-sounding speech.

However, the validation loss values, which can be found in Appendix B, suggest a gap between the training and the generalisation of the pattern to the new recordings. The total loss averaged at 2.95, which is substantially higher than the 0.69 of the training loss. Similarly, both the mel loss and the mel postnet loss averaged at 0.88, higher than their counterparts. The high values indicate that the model struggles to generalise the spectral features of the training to unseen validation samples, which could suggest some overfitting (Ren et al., 2020). The losses for pitch and energy were more pronounced at 0.63 and 0.47 respectively. These elevated values suggest the model struggled with accurately predicting expressive prosodic contours, a common challenge in emotional TTS systems (Guo et al., 2022; Ren et al., 2020). Nonetheless, the relatively stable training losses and low duration loss (0.08)

suggest that the model effectively learned key timing patterns, even if spectral and prosodic generalisation remained limited.

To evaluate the generalisation performance across different training durations, Mel Cepstral Distortion (MCD) was used at regular 100,000-step intervals . A lower value indicates a closer resemblance to the ground-truth, and thus more natural sounding synthesised speech. As shown in Figure 1, the MCD initially decreased up to 200,000 training steps, reaching a minimum of 4.63 dB. After this point MCD began to increase with a peak of 4.89 dB at 400,000 steps. Although decreasing slightly, it remained elevated at 4.83 dB for the checkpoints at 500,000, 600,000, and 700,000 steps.

Given this trend the checkpoint at 200,000 steps was chosen for further evaluation, as it showed the lowest MCD. The low spectral distortion and best generalising performance out of all the measured checkpoints suggests the model learned emotion-specific patterns from the training data without beginning to overfit. Continuing training beyond convergence could have led to diminished returns or to overfitting (Ren et al., 2020; Shen et al., 2017). The following sections evaluate this checkpoint in more detail across spectral and prosodic dimensions.

4.2 Spectral Quality: Mel-Cepstral Distortion (MCD)

In this study, Mel-Cepstral Distortion (MCD) was used to evaluate the spectral performance of both experimental setups. MCD is a widely used metric in the field of speech synthesis for evaluating the accuracy of synthesised speech compared to the original speech (Kubichek, 1993). It calculates the average frame-level distance between mel-cepstral coefficients extracted from the generated and ground-truth audio, which provides an estimate of how accurate the output sounds to the human ear. Lower MCD values suggest smaller spectral deviations and more natural-sounding synthetic speech. Typical values fall between 4 and 10 dB (Kubichek, 1993).

Table 1: Mel-Cepstral Distortion (MCD) scores calculated separately for each emotion for both setup 1 and setup 2.

Emotion	MCD-value	
	Setup 1	Setup 2
Neutral	457.42	501.45
Happy	468.65	552.35
Angry	495.62	562.99
Sad	421.11	505.30
Surprise	473.99	565.96
All	463.36	537.61

To assess how well the setups handled emotional variation, MCD was calculated separately for each of the five emotional categories in the ESD corpus, “neutral”, “happy”, “sad”, and “surprised”. The measured value is the average across all utterances in the emotional category for an overall assessment. This breakdown allowed for a more detailed comparison of spectral fidelity. The full set of results is presented in Table 1.

Setup 1, trained from scratch on the emotional subset of the ESD, consistently achieved lower MCD scores across all categories, outperforming setup 2, a fine-tuned FastSpeech 2 model pre-trained on the LJSpeech corpus. Achieving an overall MCD of 4.63 dB, compared to 5.37 dB for setup 2, setup 1 shows a stronger alignment with spectral properties of the original emotional speech. While the MCD values in both setups remain within a single decibel step, the highest levels of spectral distortion are observed in the high-intensity emotional states, “angry”, “happy”, and “surprised”. High-intensity

emotions often involve large prosodic variations and rich acoustic textures; they often tend to include rapid pitch movements, abrupt dynamic shifts, and broader spectral bandwidth. These characteristics make them more challenging to synthesise accurately as they place a higher demand on the model's capacity to replicate intricate vocal patterns. Setup 1 handled the high-intensity emotions more effectively than setup 2, this could suggest that training directly on emotionally expressive data allows the model to better capture emotional nuance despite its limited size.

In contrast, the higher MCD values in setup 2 suggest that transfer learning struggles due to domain mismatch. The LJSpeech corpus used for the pre-training consists entirely of neutral speech by a single female speaker, making its acoustic profile different from the emotionally expressive ESD used during fine-tuning. The pre-trained model may have retained the spectral and prosodic patterns which were optimised for neutral speech. This pre-training bias (Latif et al., 2020) could have made it less adapting to the broader range of expression in emotional speech, which can result in "spectral blurring" or oversmoothing, where the model is unable to fully capture the rapid vocal changes and dynamic shifts needed to convincingly reproduce emotions such as anger and surprise.

One particularly telling observation is that the lowest MCD in setup 1 was found for sad, measured at 4.21 dB. Sad speech is generally characterised by lower pitch, slower speech rate, and more stable acoustic features. Compared to the mid-range values for pitch, intensity, and tempo seen in neutral speech (Pell et al., 2009), sad speech is acoustically simpler, placing fewer demands on the model's ability to capture dynamic prosodic variation. This suggests that FastSpeech 2 can effectively synthesise emotions with lower prosodic variability. However, in setup 2 the MCD values for neutral and sad speech were nearly identical, differing only by 0.04 dB. This could suggest that the pre-training bias toward neutral speech may have benefitted the model when handling sad speech, likely due to similarities in pitch stability and overall spectral simplicity between the two. Sad speech, however, still performed poorer than neutral speech, showing that the influence of pre-training effects even the lower-complexity emotions.

Although MCD offers important insights into the spectral fidelity of synthesised speech, it alone is not sufficient to evaluate emotional expressiveness. Pitch, duration, and intensity are important factors (Banse & Scherer, 1996) and contribute to the perception of emotion in speech. The next sections give insight into the prosodic performance of both models to assess how effectively they capture and reproduce emotional speech.

4.3 Emotional Prosody Patterns

Prosodic features, such as pitch (F0), duration, and intensity, function as reliable acoustic markers of emotional expression. Different emotions are consistently associated with distinct patterns across these parameters (Juslin & Laukka, 2003). Table 2 summarises the average values across emotion categories, which are compared to provide insight into how effectively emotional expressiveness is conveyed by the two model setups relative to human speech.

4.3.1 Ground-Truth Patterns

The emotional speech from the ESD, which serves as the ground-truth, shows prosodic variation that partly reflects patterns observed in previous studies on emotional speech. "Angry", "happy", and "surprised" are typically classed as high-intensity emotions, their speech is typically characterised by

Table 2: Average prosodic values (pitch, duration, intensity) per emotion category for ground truth, setup 1, and setup 2.

Emotion	Ground-truth			Setup1			Setup2		
	Duration (s)	Pitch (Hz)	Intensity (dB)	Duration (s)	Pitch (Hz)	Intensity (dB)	Duration (s)	Pitch (Hz)	Intensity (dB)
Neutral	1.99	210	63.34	1.93	202	69.65	1.87	203	45.82
Happy	2.22	258	59.93	2.15	250	64.85	2.04	246	67.53
Angry	1.91	264	63.40	1.82	252	67.84	1.76	253	69.67
Sad	2.41	230	66.90	2.33	215	70.88	2.18	218	71.25
Surprise	1.98	317	63.19	1.91	316	67.91	1.82	305	69.24
All	2.10	256	63.35	2.03	247	68.23	1.93	245	69.67

higher pitch, greater intensity, and faster speech rates. Sad and neutral speech, in contrast, tend to be slower and lower in pitch (Goudbeek & Scherer, 2010; Mozziconacci, 2002).

The highest pitch in the ground-truth data measured at 317 Hz for surprised speech, followed by angry and happy at respectively 264 and 258 Hz. These values were all notably higher than neutral speech (210 Hz). It aligns with what is expected for these emotions and indicate strong pitch-based expressivity (Juslin & Laukka, 2003; Scherer, 2003). Sad speech had, at 230 Hz, a lower pitch than the other emotions but still higher than neutral. Although sad speech is typically described as low-arousal and therefore expected to show a lower pitch, several studies have noted that acted or read sad speech can involve a slightly raised F0. This is often attributed to increased vocal tension or a breathy voice quality (Banse & Scherer, 1996; Mozziconacci, 2002).

Sad speech unexpectedly showed the highest average intensity at 67 dB, contradicting previous the notion where sadness is generally associated with reduced vocal energy (Juslin & Laukka, 2003). Happy speech, by contrast, showed the lowest intensity at 60 dB. The other emotional states, namely angry, neutral, and surprised, had a nearly identical intensity average at 63 dB. This could suggest speaker-specific variation, such as expressing happiness more like calm contentment rather than excited joy (Goudbeek & Scherer, 2010; Juslin & Laukka, 2003) and anger and surprise with more tense vocal quality rather than loudness (Scherer, 1986). Additionally, recording-related factors, such as , led to a more subdued delivery style by supressing amplitude differences (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Livingstone & Russo, 2018).

In terms of duration, sad speech had the longest average duration at 2.41 seconds, reflecting a slower speech rate and elongated utterances often associated with low-arousal emotional states. Angry and surprised speech, with durations of 1.91 and 1.98 seconds respectively, were also consistent with prior studies, showing a faster pace compared to neutral (1.99 seconds) (Goudbeek & Scherer, 2010). Happy speech, however, a duration shorter than expected of 2.22 seconds. Similar to the intensity result, this could be reflected by either speaker-specific variation or the influence from controlled studio settings (Burkhardt et al., 2005; Livingstone & Russo, 2018).

Understanding the prosodic patterns of the ground-truth data provides a reference point against which the models' performance can be judged. Some ground-truth patterns match what has been found in previous studies. Such as higher pitch for high-intensity states. Others, particularly the intensity, do not. These nuances should be kept in mind when evaluating how well the model replicates human emotional prosody.

4.3.2 Setup 1: Training from Scratch

Setup 1's output broadly mirrored the prosodic patterns observed in the ground-truth. The average pitch values per emotion were relatively consistent to the ground-truth. Surprised speech was measured to have the highest pitch at 316 Hz, closely matching the ground-truth value (317 Hz). Angry and happy speech followed at 252 Hz and 250 Hz respectively, again aligning with the expectations for these high-intensity emotions (Juslin & Laukka, 2003; Scherer, 2003). Neutral and sad speech had the lowest pitch values, at 202 Hz and 215 Hz respectively, consistent with what is known about low arousal emotions (Goudbeek & Scherer, 2010).

The produced output of setup 1 generally had louder speech compared to the ground-truth, with values ranging from 65 dB for happy speech to 71 dB for sad. The order in which the categories increase intensity has changed. Sad remains the most intense, followed by neutral, surprised, and angry speech (resp. 70 dB, 68 dB, and 68 dB). Happy showed the lowest intensity, again. The range of the intensity of the emotions ranged in the middle has widened in setup 1 compared to the ground-truth, going from 0.2 dB to 1.8 dB. A possible reason for this increase is an overemphasis of prosodic distinctions by the model to enhance the differences between the given categories. This effect is often seen in expressive TTS when trained on limited data, the model will rely more on surface-level cues, like intensity, to represent the emotion (Auzawa et al., 2018; Kim et al., 2022).

Duration patterns were reasonably well preserved. Sad speech had the longest average duration with 2.33 seconds. Happy speech followed with 2.15 seconds. Neutral, surprised, and angry speech all stayed under 2 seconds. This suggests the model successfully reproduced the speech rates associated with their respective emotional category based on the ground-truth recordings. Duration is often more accurately reproduced in TTS systems compared to other prosodic features. It is easier to model and is less variable across emotional states (Ren et al., 2020; Shen et al., 2017; Stanton et al., 2018).

The findings indicate that setup 1 was able to learn and replicate key prosodic distinctions among the emotional categories based on the ground-truth recordings. This suggests that training directly on emotional speech enables the model to replicate natural prosodic variation effectively, especially in duration and pitch. However, the elevated intensity values may reflect a modelling bias towards intensity or limitations in the model's ability to calibrate amplitude levels precisely.

4.3.3 Setup 2: Fine-Tuning from LJSpeech

The prosodic output of setup 2, which involve fine-tuning a pre-trained FastSpeech 2 model, showed reduced expressiveness across all features. Overall the pitch measured around 11 Hz lower than compared to the ground-truth data, with surprised speech showing the highest F0 at 305 Hz. This was followed by angry at 253 Hz, happy at 246 Hz, and sad at 218 Hz. The pitch for neutral speech was most similar in value to the original data at 203 Hz. While the 11 Hz difference is subtle and unlikely to be perceived by the human ear (Pell et al., 2009), it could suggest underlying limitations in the model's ability. The LJSpeech corpus, consisting of solely neutral speech with relatively flat intonation (Latif et al., 2020; Shen et al., 2017; Stanton et al., 2018), could have introduced pre-training bias. This bias can restrict the model's capacity to learn higher pitch contours, even after fine-tuning (Latif et al., 2020). The relative pitch distribution across the emotional categories remained largely consistent to the ground-truth, only at a lower baseline.

Table 3: Mean Absolute Error (MAE) values for duration, pitch, and intensity across emotional states for ground truth, setup 1, and setup 2.

Emotion	MAE1			MAE2		
	Duration (s)	Pitch (Hz)	Intensity (dB)	Duration (s)	Pitch (Hz)	Intensity (dB)
Neutral	0.13	11	6.31	0.22	15	7.33
Happy	0.17	14	5.01	0.29	24	7.61
Angry	0.14	17	4.56	0.21	25	6.31
Sad	0.16	17	4.06	0.27	20	4.41
Surprise	0.15	18	4.87	0.56	38	6.36
All	0.15	15	4.96	0.31	24	6.40

Duration values in setup 2 followed the ordering as seen in the ground-truth, with sad taking the longest and angry the quickest, 2.18 and 1.76 seconds respectively. The overall durations are lower than the original ground-truth data, suggesting a compressed temporal structure likely influenced by the pre-training on the speech pace of the LJSpeech data (Goudbeek & Scherer, 2010).

The biggest difference can be seen in intensity. Sad and angry speech have the highest intensity, similar to the ground-truth, at 71 dB and 70 dB respectively. This is followed by surprised and happy speech at 69 dB and 68 dB. While these values rank higher than in the original data, their intensity remains relatively close to the ground-truth. Neutral speech, however, measured at a notably low 46 dB. This large gap may indicate a domain mismatch. The model could be reverting to the lower intensity levels typical of the LJSpeech corpus, which lacks emotional variation and is recorded in a subdued manner (Latif et al., 2020; Shen et al., 2017; Stanton et al., 2018).

While pitch was fairly accurate produced, limitations could be seen in duration and loudness. The influence of the neutral LJSpeech pre-training likely affected the model's capacity to fully express prosodic variation. This can be seen by the lowered intensity range and shortened durations, suggesting that the model retained aspects of the neutral prosodic baseline from the pre-raining data (Burkhardt et al., 2005; Latif et al., 2020).

4.4 Prosodic Accuracy: MAE for Pitch, Duration, and Intensity

For a more detailed picture of how accurately each model reproduces expressive speech the Mean Absolute Error (MAE) was used. MAE calculates the average absolute difference between the predicted and ground-truth values for pitch, intensity, and duration. These values can be seen in Table 3 separated by emotional category for both setup 1 and setup 2. A lower MAE score indicates a closer match to the original recordings and better generalisation of emotion-specific prosodic characteristics.

4.4.1 Overall Prosodic Accuracy

Setup 1 showed lower MAE scores across all measured prosodic features. The pitch was over 9 Hz lower in setup 1 with 15.35 Hz compared to 24.28 setup 2. Pitch variation is a key factor in conveying emotional valence and arousal (Mozziconacci, 2002; Schröder, 2001), making this a noticeable difference in expressive speech. Similarly, duration error for setup 2 is twice as much as for setup 1, 0.31 and 0.15 seconds respectively. The lower pitch and duration error suggest that the model captured

both the contour and the timing of emotional speech more accurately in setup 1. Subtle prosodic variation contributes significantly to perceived emotional quality (Juslin & Laukka, 2003; Scherer, 2003), as humans can detect temporal deviations as small as 50-100 milliseconds (Nakatani, O'Connor, & Aston, 1981; Rosen, 1992), capturing it adequately is essential in expressive speech synthesis.

Setup 1 also outperformed setup 2 in intensity accuracy, with an overall MAE of 4.96 dB compared to 6.40 dB. The smaller intensity error suggest that setup 1 more effectively matched the loudness of each emotion to what is learned from the ground-truth compared to setup 2. In contrast, the higher MAE values in setup 2 likely reflect the influence from its neutral pre-training, which may have biased the model towards more monotonic and restrained prosodic patterns (Latif et al., 2020; Shen et al., 2017; Stanton et al., 2018), limiting its adaptability to the emotional variability present in the ESD training data.

4.4.2 Emotion-Specific Prosodic Accuracy

When broken down by emotion, setup 1 consistently showed lower MAE values than setup 2. In both setups happy, sad, and surprised speech showed the highest duration errors, though in different orders. This could indicate that these emotions, which are often characterised by more complex or variable speech timing (Goudbeek & Scherer, 2010), were generally harder to synthesise accurately in both training conditions. However, sad speech does not fit these characteristics, as it is typically described as having slower and more regular timing patterns with fewer dynamic variation (Cowie & Cornelius, 2003; Scherer, 1986). Its high error may instead reflect the model's difficulty in maintaining consistent slow timing. The difference in ranking order across the setups may reflect how each training strategy impacts the model's sensitivity to timing cues in emotion-specific ways. The biggest difference between the setups can be seen with surprised speech, in both duration and pitch. This suggests that surprised speech posed the greatest challenge to setup 2, the fine-tuned model, which aligns with previous reports that high-intensity emotions are more difficult to reproduce due to their rapid prosodic modulation (Goudbeek & Scherer, 2010; Mozziconacci, 2002). Neutral and angry speech had low duration error compared to the other emotions in both setups, 0.13 and 0.14 respectively for setup 1, and 0.22 and 0.21 for setup 2. These emotion were generally easier to reproduce as they have more predictable or moderate timing patterns (Goudbeek & Scherer, 2010; Scherer, 1986).

The pitch MAE indicates that high-intensity emotions, particularly surprise and anger, are the hardest to model accurately. Setup 2 shows greater difficulty than setup 1, with a 25 Hz measured for angry speech and 38 Hz for surprised, reflecting the challenge of capturing prosodic variations typical of these emotions (Goudbeek & Scherer, 2010; Mozziconacci, 2002; Schröder, 2001). While happy speech shows a lower MAE in setup 1, as its prosodic patterns are moderate variable but remain relatively regular, making it easier to reproduce than surprise or angry speech. However, in setup 2 happy speech has a higher pitch error (24 Hz), on par with the other high-intensity emotions, this is likely due to the model's difficulty in capturing the expressive variability after pre-training on neutral data. Sad speech, despite being low-intensity and slower, still has a high MAE value, especially in setup 2 (20 Hz). This suggest that subtle prosodic variations are challenging to capture when fine-tuning from neutral speech.

In terms of intensity setup 1 again showed lower MAE values across all emotions. The lowest value is observed for sad speech, with 4.06 dB, which is consistent with what is expected from low-intensity emotions (Cowie & Cornelius, 2003; Scherer, 1986). Angry speech also showed a low intensity error

(4.56 dB), suggesting that learning from scratch enables the model to capture the high vocal energy of this emotion. In contrast, setup 2 showed higher MAE values for most emotions, particularly for happy and surprised speech, 7.61 dB and 6.36 dB respectively. This could indicate that the neutral pre-training makes the model either exaggerate or downplay sudden intensity fluctuations. Interestingly, even neutral speech shows higher intensity error in setup 2 (7.33 dB) compared to setup 1 (6.31), despite its neutral pre-training. This could further suggest that the expressive training in setup 1 not only enhances emotional variation but also supports the baseline loudness patterns.

4.5 Implications for Emotional TTS

The findings of this study offer insights that could guide the development of emotionally expressive TTS systems. Firstly, training FastSpeech 2 on emotional data from scratch (setup 1) proved more efficient than fine-tuning from a neutral pre-trained model. Despite the small size of the dataset, the model successfully learned prosodic patterns corresponding to different emotions, resulting in better spectral fidelity and more accurate prosodic patterns. These outcomes align with prior work that highlight emotion-disentangled and prosody-aware training objectives for expressive synthesis (Auzawa et al., 2018; Hsu et al., 2019).

Secondly, fine-tuning alone, without explicit emotion representations or adaptation mechanisms for emotion, may limit transfer learning. Setup 2 maintained the neutral prosodic patterns from its pre-trained configuration, which led to less dynamic prosody with limited pitch and intensity variation, despite fine-tuning on emotional speech. This confirms that transfer learning between domains with divergent prosodic characteristics is not necessarily effective without techniques such as emotion embeddings, attention-based adaptation, or emotion classifiers are added (Stanton et al., 2018; S. Zhang et al., 2025).

Lastly, these insights could inform more effective approaches for low-resource languages and voices. When data are limited training a model from scratch on a small emotional dataset can outperform fine-tuning from a large neutral dataset. For languages where fully labelled emotional data are not available, training small models from scratch may appear as the most practical starting point.

4.6 Comparison with Previous Research

This work demonstrates that FastSpeech 2, when trained on emotional data, can capture a substantial amount of emotion-specific variation in pitch, duration, and intensity without explicit emotion embeddings. While many recent systems rely on external mechanisms such as encoders or style tokens to control emotional variation (Shen et al., 2017; Stanton et al., 2018), the findings in this study indicate that FastSpeech 2 internal variance predictors are capable of implicitly learning prosodic patterns when the training dataset contains sufficient emotional content. This aligns with prior findings that reinforce the view that neural TTS can learn expressive traits from prosody alone (Ren et al., 2020; Wu et al., 2022). However, building upon the prior studies, a detailed evaluation across multiple emotional categories and prosodic metrics are given here.

The prosodic accuracy observed in setup 1 is consistent with results from other low-resource, single-speaker studies (Zhou et al., 2021). In that study high-intensity emotions such as “surprise” and “anger” showed increased pitch and decreased duration, while “sad” speech was flatter and had prolonged duration. The same is seen in setup 1. That these patterns appeared even without explicit

emotional guidance indicate that variance predictors are capable of modelling emotional prosody effectively when exposed to suitable data. Setup 2, however, underperformed across all metrics, despite success in transfer learning across domains such as speaker adaptation or accent conversion (Udagawa et al., 2022; Valle, Li, Prenger, & Catanzaro, 2019). A possible explanation is that emotional prosody differs from speaker identity or intelligibility; emotional expression involves more subtle prosodic variations, which can diminish dynamic variations during fine-tuning from neutral speech. Pre-training on neutral data restricted the expressive range of the fine-tuned model, mirroring earlier studies on the need to disentangle speaker and prosodic representations (Auzawa et al., 2018; Hsu et al., 2019).

Recent work have introduced approaches such as multimodal and reference-guided emotion control (e.g. GST-Tacotron, Stanton et al. (2018)) that allow zero-shot emotion transfer. Speech synthesis systems are then able to use the information learned from emotion by generalising to generate speech in unseen emotional style, without the need for explicit training examples for that emotion. While highly flexible, these methods typically require an extensive and well-annotated emotional dataset and are prone to suffer from training noise or entanglement. This study shows that FastSpeech 2, trained under the right conditions, can act as a scalable and lightweight option for expressive TTS. This is particularly the case when spectral quality and prosodic fidelity are prioritised over fine-grained controllability.

4.7 Limitations and Future Directions

This study on the emotional expressiveness of FastSpeech 2 has several limitations that warrant consideration. It was chosen to use only objective metrics, like the Mel-Cepstral Distortion (MCD) and Mean Absolute Error (MAE), to assess expressive accuracy. These are metrics that offer reproducibility and interpretability (Cowie et al., 2001; Kubichek, 1993), however, they are limited in their ability to represent the nuanced emotional significance that can be interpreted through human perception. Due to the absence of subjective evaluation (e.g. listener ratings or preference tests) the naturalness and affective clarity of the synthesised speech could not be effectively determined. This constraint was primarily due to resource limitations and the lack of sufficient participant availability.

Another limitation is the size of the dataset. The emotional speech subset from the ESD corpus (Zhou et al., 2021), even though comprising of five emotions, remains relatively small in size compared to datasets typically used in TTS research. This smaller amount to train the model on in setup 1 increased the risk of overfitting. To prevent this from happening regularisation strategies, such as increased dropout, weight decay, and learning rate adjustments, were employed. These measures, however, may not fully eliminate generalisation issues. Furthermore, while the fine-tuning in setup 2 aimed to make use of the pre-trained prosodic representations from the neutral LJSpeech dataset, the mismatch in speaker identity and recording conditions between LJSpeech and ESD may have limited how effective transfer learning could be, especially in replicating the fine-grained emotional variation.

In addition, only five emotional categories were used in this study, namely “neutral”, “happy”, “angry”, “sad”, and “surprised”. These are basic and categorical emotions, neglecting more complex, nuanced or culturally specific emotions (e.g. jealousy, pride, calmness). Adding complex emotions tends to broaden the generalisability of the results (Banse & Scherer, 1996). Similarly, only the average of the prosodic features were examined more closely. There was no accounting for dynamic prosodic contours or temporal variation, other than mentioning the range, which may play crucial

roles in emotional expressivity (Cowie et al., 2001).

Several directions for future work can be identified. Incorporating subjective listening evaluations, in combination with objective evaluations, could greatly enrich the analysis. Acoustic metrics could then be aligned with human perception (Schröder, 2009). Secondly, creating a bigger dataset, through either data augmentation, synthetic resampling, or inclusion of emotional data from multiple speakers. This could improve model robustness and could give insights into speaker-dependent and speaker-independent emotional expressiveness (Wu et al., 2022). Third, knowing how FastSpeech 2 generates emotions without embeddings or style tokens opens the door to integrate them into the architecture and explore how they advance the model. Finally, further work could focus on the temporal evolution of prosody across utterances, which static averages may hide.

5 Conclusion

In this thesis it was examined whether FastSpeech 2 can convey emotional prosody in synthesised speech without explicit emotion embeddings. Two strategies were compared: setup 1 trained from scratch on a small emotional dataset and setup 2 fine-tuned a neutral pre-trained model on the same data. Findings show that FastSpeech 2 can learn emotional prosody directly from the dataset alone, with training from scratch achieving more accurate prosodic patterns than fine-tuning a neutral model. Analysis of spectral fidelity, measured using MCD, indicated that setup 1 consistently achieved lower distortion than setup 2 for all emotions. Low-arousal emotions, such as sadness, had the smallest spectral deviation, whereas the high-arousal emotions, namely happiness, anger and surprise, produced higher MCD scores reflecting more dynamic acoustic patterns. The elevated MCD in setup 2 suggests that transfer learning from neutral speech introduces a pre-training bias that reduce the model's ability to capture emotion-specific acoustic variation.

Evaluation of prosodic accuracy, using MAE, for pitch, duration, and intensity, further supported setup 1. Pitch and duration were more accurately reproduced, while setup 2 showed flattened prosodic contours, especially for high-arousal emotions. The reduced prosodic variability seen in setup 2 suggest than fine-tuning from neutral speech constraints pitch and duration variation, reducing the prosodic dynamics needed for emotional expression. The model trained from scratch occasionally overstated the intensity levels, suggesting that the model leverages prominent features when trained on small datasets. FastSpeech 2 can reproduce emotion-specific without explicit labels, as long as he training data captured the adequate prosodic diversity. Although pre-training is often used in low-resource tasks, it may suppress expressive variation when the initial dataset is neutral. Alignment between the training data and the emotional target task is more important than dataset size,; smaller and expressive datasets can outperform fine-tuning from too large neutral models.

These results have practical implications. When collecting data, emphasis should be placed on the variability of expression rather than dataset size alone. Simple model can capture emotion-specific prosody without requiring style tokens or reference encoders. These could be better used for fine-grained control or zero-shot synthesis. Combining small emotional datasets with lightweight adaption methods offers greater flexibility without introducing pre-training biases. This study also had its limitations, namely its reliance on objective metrics and a restricted set of discrete emotions. Future work should combine human perceptual evaluations and more nuanced emotional categories to measure both naturalness and affective clarity. Nevertheless, the results confirm that FastSpeech 2 can reproduce emotion-specific prosody directly from data when it covers the full spectrum of prosodic variation.

References

Auzawa, K., Iwasawa, Y., & Matsuo, Y. (2018). Expressive speech synthesis via modeling expressions with variational autoencoder. In *Interspeech 2018* (pp. 3067–3071). doi: 10.21437/Interspeech.2018-1113

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*. doi: 10.1037/0022-3514.70.3.614

Barakat, H., Turk, O., & Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources. *EURASIP journal on Audio Speech and Music Processing*, *2024*. doi: 10.1186/s13636-024-00329-7

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech. In *Interspeech 2005* (pp. 1517–1520). doi: 10.21437/Interspeech.2005-446

Campbell, N., & Mokhtari, P. (2003). Voice quality: The 4th prosodic dimension..

Chang, H.-S., Lee, C.-Y., Wang, X., Young, S.-T., Li, C.-H., & Chu, W.-C. (2023). Emotional tones of voice affect the acoustics and perception of mandarin tones. *PLOS ONE*, *18*. doi: 10.1371/journal.pone.0283635

Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., & Lee, S.-W. (2024). Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech 2024* (pp. 1810–1814). doi: 10.21437/Interspeech.2024-398

Cornille, T., Wang, F., & Bekker, J. (2022). Interactive multi-level prosody control for expressive speech synthesis. In (pp. 8312–8316). doi: 10.1109/ICASSP43922.2022.9746654

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

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, *18*(1), 32–80.

Cui, C., Ren, Y., Liu, J., Chen, F., Huang, R., Lei, M., & Zhao, Z. (2021). Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model. In *Interspeech 2021* (pp. 2766–2770). doi: 10.21437/Interspeech.2021-1148

Cui, Y., Wang, X., Zhao, Z., Zhou, W., & Chen, H. (2023). Casein: Cascading explicit and implicit control for fine-grained emotion intensity regulation. In *Interspeech 2023* (pp. 4813–4817). doi: 10.21437/Interspeech.2023-843

Diatlova, D., & Shutov, V. (2023). Emospeech: guiding fastspeech2 towards emotional text to speech. In *12th isca speech synthesis workshop (ssw2023)* (pp. 106–112). doi: 10.21437/SSW.2023-17

Ekman, P. (1992). An argument for basic emotions. *Cognition Emotion*, *6*, 169–200. doi: 10.1080/02699939208411068

Eyben, F., Scherer, K. R., Schuller, B., Sundberg, J., Andre, E., Busso, C., ... Truong, K. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*, 1–1. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). opensmile: The munich versatile and fast open-source audio feature extractor. In (pp. 1459–1462). doi: 10.1145/1873951.1874246

Galdino, J., Matos, A., Svartman, F., & Aluisio, S. (2025). The evaluation of prosody in speech synthesis: a systematic review. *Journal of the Brazilian Computer Society*, *31*, 466–487. doi: 10.5753/jbcs.2025.5468

Gobl, C., & Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and at titude. *Speech Communication*, *40*, 189–212. doi: 10.1016/S0167-6393(02)00082-1

Goudbeek, M., & Scherer, K. R. (2010). Beyond arousal: Valence and potency/control cues in the

vocal expression of emotion. *The journal of the Acoustical Society of America*, *128*, 1322–36. doi: 10.1121/1.3466853

Guo, Y., Du, C., Chen, X., & Yu, K. (2022). Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. doi: 10.48550/arXiv.2211.09496

He, J., Gong, C., Wang, L., Jin, D., Wang, X., Xu, J., & Dang, J. (2022). Improve emotional speech synthesis quality by learning explicit and implicit representations with semi-supervised training. In *Interspeech 2022* (pp. 5538–5542). doi: 10.21437/Interspeech.2022-11336

Hirst, D., & di Cristo, A. (1998). *Intonation systems: A survey of twenty languages* (Vol. 76). doi: 10.2307/417674

Hsu, W.-N., Zhang, Y., Weiss, R., Chung, Y.-A., Wu, Y., & Glass, J. (2019). Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In (pp. 5901–5905). doi: 10.1109/ICASSP.2019.8683561

Ikeda, M., & Markov, K. (2024). Fastspeech2 based japanese emotional speech synthesis. In (pp. 1–5). doi: 10.1109/IS61756.2024.10705252

Im, C.-B., Lee, S.-H., Kim, S.-B., & Lee, S.-W. (2022). Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. In (pp. 6317–6321). doi: 10.1109/ICASSP43922.2022.9747098

Inoue, S., Zhou, K., Wang, S., & Li, H. (2024). *Hierarchical emotion prediction and control in text-to-speech synthesis.* doi: 10.48550/arXiv.2405.09171

Izard, C. E. (1993). Four systems for emotion activation: cognitive and noncognitive processes. *Psychological Review*, *100*, 68–90. doi: 10.1037//0033-295X.100.1.68

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

Kim, C., Um, S., Yoon, H., & Kang, H.-G. (2022). Fluenttts: Text-dependent fine-grained style control for multi-style tts. In *Interspeech 2022* (pp. 4561–4565). doi: 10.21437/Interspeech.2022-988

Kong, J., Jaehyeon, K., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. doi: 10.48550/arXiv.2010.05646

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In (pp. 125–128). doi: 10.1109/PACRIM.1993.407206

Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., … Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis.

Kögel, F., Nguyen, B., & Cardinaux, F. (2023). Towards robust fastspeech 2 by modelling residual multimodality. In *Interspeech 2023* (pp. 4309–4313). doi: 10.21437/Interspeech.2023-879

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Larrouy-Maestri, P., Poeppel, D., & Pell, M. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*, *20*. doi: 10.1177/17456916231217722

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. doi: 10.48550/arXiv.2001.00378

Laukka, P., & Elfenbein, H. (2020). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review*, *13*. doi: 10.1177/1754073919897295

Lee, Y., Rabiee, A., & Lee, S. Y. (2017). Emotional end-to-end neural speech synthesizer.

doi: 10.48550/arXiv.1711.05447

Lenglet, M., Perrotin, O., & Bailly, G. (2023). Local style tokens: Fine-grained prosodic representations for tts expressive control. In *12th isca speech synthesis workshop (ssw2023)* (pp. 120–126). doi: 10.21437/SSW.2023-19

Li, Y., Zhu, X., Lei, Y., Li, H., Liu, J., Xie, D., & Xie, L. (2023). Zero-shot emotion transfer for cross-lingual speech synthesis. In (pp. 1–8). doi: 10.1109/ASRU57964.2023.10389638

Liu, R., Yifan, H., Ren, Y., Yin, X., & Li, H. (2024). Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 18698–18706. doi: 10.1609/aaai.v38i17.29833

Livingstone, S., & Russo, F. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, *13*. doi: 10.1371/journal.pone.0196391

Mozziconacci, S. (2002). Prosody and emotions. In *Speech prosody 2002* (pp. 1–9). doi: 10.21437/SpeechProsody.2002-1

Nakatani, L., O'Connor, K., & Aston, C. (1981). Prosodic aspects of american english speech rhythm. *Phonetica*, *38*, 84–105. doi: 10.1159/000260016

Pell, M., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. (2009). Factors in the recognition of vocally expressed emotions : A comparison of four languages. *Journal of Phonetics, v.37, 417-435 (2009)*, *37*. doi: 10.1016/j.wocn.2009.07.005

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: fast, robust and controllable text to speech. *Neural Information Processing Systems*, *32*, 3165–3174.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *336*, 367–73. doi: 10.1098/rstb.1992.0070

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161–1178. doi: 10.1037/h0077714

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*, 143–165. doi: 10.1037/0033-2909.99.2.143

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*, 227–256.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*. doi: 10.1080/02699930902928969

Schröder, M. (2001). Emotional speech synthesis: a review. In *7th european conference on speech communication and technology (eurospeech 2001)* (pp. 561–564). doi: 10.21437/Eurospeech.2001-150

Schröder, M. (2004). Emotional speech synthesis.

Schröder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In (pp. 111–126). doi: 10.1007/978-1-84800-306-4$_7$

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*, 1062–1087. doi: 10.1016/j.specom.2011.01.011

Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2017). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.
doi: 10.48550/arXiv.1712.05884

Shi, H., Wang, J., Zhang, X., Cheng, N., Yu, J., & Xiao, J. (2024). Rset: Remapping-based sorting

method for emotion transfer speech synthesis. In (pp. 90–104). doi: 10.1007/978-981-97-7232-2_7

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

Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., ... Saurous, R. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. doi: 10.48550/arXiv.1803.09017

Tang, H., Zhang, X., Cheng, N., Xiao, J., & Wang, J. (2024). Ed-tts: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis.. doi: 10.1109/ICASSP48485.2024.10446467

Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023a). Emomix: Emotion mixing via diffusion models for emotional speech synthesis. In *Interspeech 2023* (pp. 12–16). doi: 10.21437/Interspeech.2023-1317

Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023b). Qi-tts: Questioning intonation control for emotional speech synthesis. In (pp. 1–5). doi: 10.1109/ICASSP49357.2023.10095623

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Themistocleous, C. (2025). Linguistic and emotional prosody: A systematic review and ale meta-analysis. *Neuroscience Biobehavioral Reviews*, *175*. doi: 10.1016/j.neubiorev.2025.106210

Toda, T., Black, A., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, *15*, 2222–2235. doi: 10.1109/TASL.2007.907344

Udagawa, K., Saito, Y., & Saruwatari, H. (2022). Human-in-the-loop speaker adaptation for dnn-based multi-speaker tts. In *Interspeech 2022* (pp. 2968–2972). doi: 10.21437/Interspeech.2022-257

Valle, R., Li, J., Prenger, R., & Catanzaro, B. (2019). Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. doi: 10.48550/arXiv.1910.11997

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, *7*, 1–11. doi: 10.1038/s41562-022-01505-5

Wang, X., Fang, R., & Ding, H. (2024). Gender differences in acoustic-perceptual mapping of emotional prosody in mandarin speech. *Corpus-based Studies across Humanities*, *2*, 235–264. doi: 10.1515/csh-2024-0025

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. In *Interspeech 2017* (pp. 4006–4010). doi: 10.21437/Interspeech.2017-1452

Weide, R. (1998). The cmu pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict*.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, *30*. doi: 10.3354/cr030079

Wu, Y., Wang, X., Zhang, S., He, L., Song, R., & Nie, J.-Y. (2022). Self-supervised context-aware style representation for expressive speech synthesis. In *Interspeech 2022* (pp. 5503–5507). doi: 10.21437/Interspeech.2022-686

Xu, Y. (2019). Prosody, tone, and intonation. In *The routledge handbook of phonetics* (pp. 314–356).

Routledge.

Xue, J., Deng, Y., Han, Y., Li, Y., Sun, J., & Liang, J. (2022). Ecapa-tdnn for multi-speaker text-to-speech synthesis. In (pp. 230–234). doi: 10.1109/ISCSLP57327.2022.10037956

Zhang, J., Wushouer, M., Tuerhong, G., & Wang, H. (2023). Semi-supervised learning for robust emotional speech synthesis with limited data. *Applied Sciences*, *13*.

Zhang, S., Mehrish, A., Li, Y., & Poria, S. (2025). Proemo: Prompt-driven text-to-speech synthesis based on emotion and intensity control.
doi: 10.48550/arXiv.2501.06276

Zhou, K., Sisman, B., Liu, R., & Li, H. (2021). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, *137*, 1–18. doi: 10.1016/j.specom.2021.11.006

Zhu, X., Lei, Y., Li, T., Zhang, Y., Zhou, H., Lu, H., & Xie, L. (2024). Metts: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *PP*, 1–13. doi: 10.1109/TASLP.2024.3363444

Appendices

A Training and validation loss curves for Setup 1

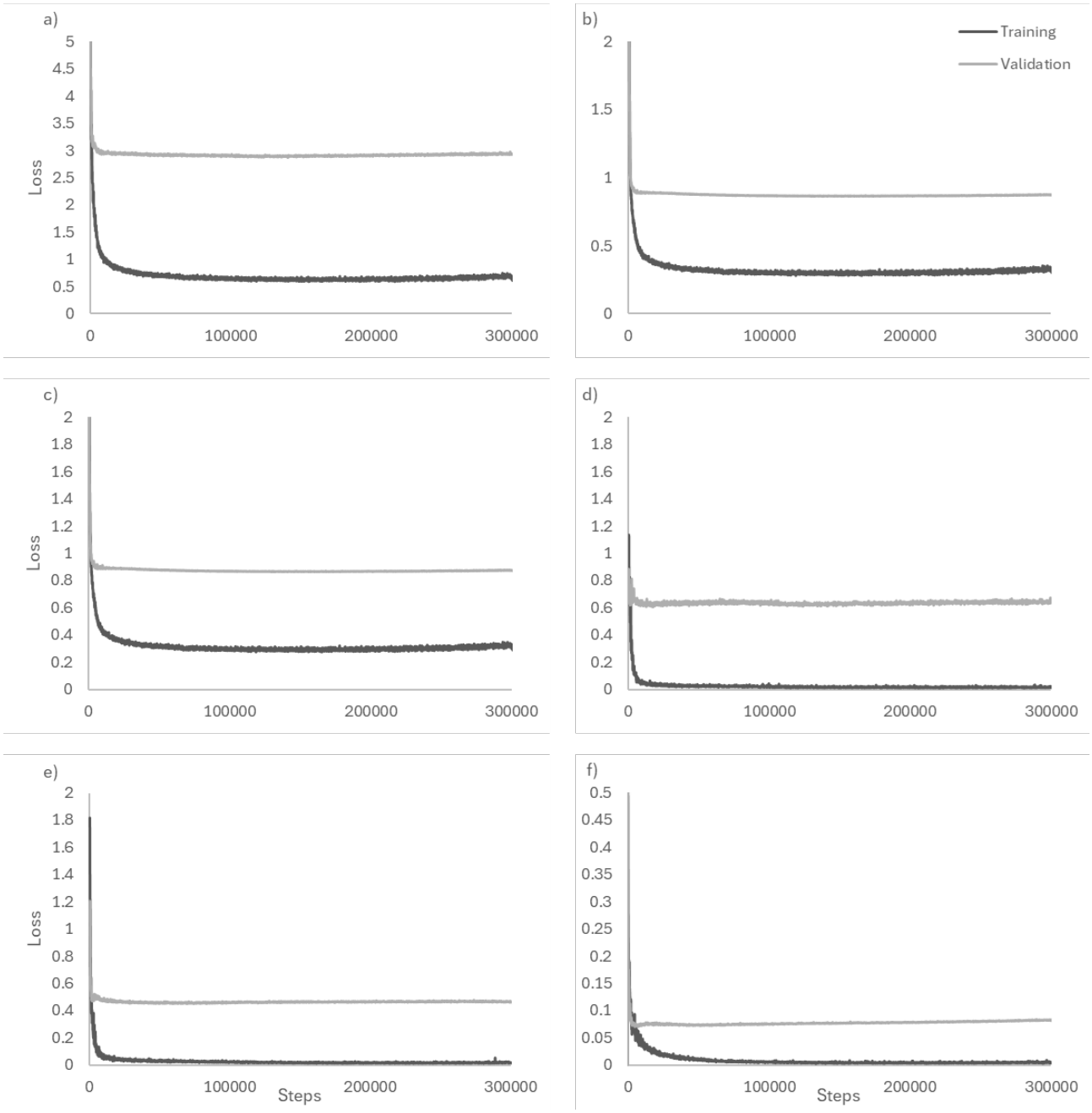

Figure 2: Overall loss (a), mel loss (b), mel post-net loss (c), pitch loss (d), energy loss (e), and duration loss (f) across training steps. For each panel, the training and validation loss curves illustrate how the model converges over time and reveal differences in learning stability across the various sub-loss components.

B Training and validation loss curves for Setup 2

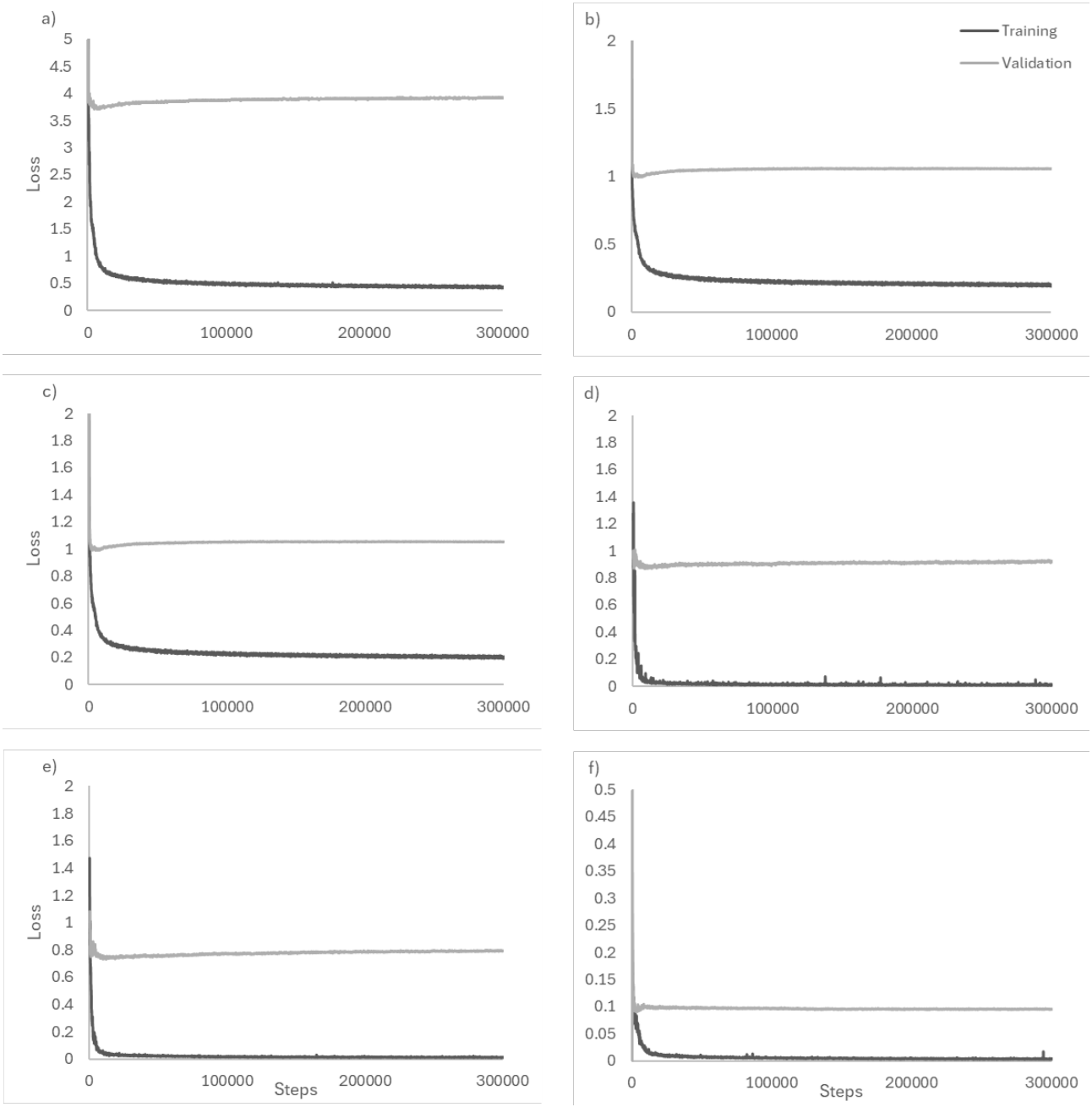

Figure 3: Overall loss (a), mel loss (b), mel post-net loss (c), pitch loss (d), energy loss (e), and duration loss (f) across training steps. For each panel, the training and validation loss curves illustrate how the model converges over time and reveal differences in learning stability across the various sub-loss components.