Enhancing Prosody Transfer in Speech Synthesis with Prosodically-Aligned References

Lin.Liu

Master of Science
School of Informatics
University of Edinburgh
2024

Abstract

Prosody transfer in speech synthesis plays a crucial role in producing natural and expressive speech by mimicking the prosody of reference speech. Traditional methods that rely on ground truth references during training often perform well but struggle to generalize during inference when the reference differs from the target, leading to degraded quality and speaker leakage issue. To address these challenges, we introduce a method that leverages prosodically-aligned speech as references during training, generated through unit selection. This approach ensures more consistent performance across varied reference types, preserves the target speaker's timbre, and achieves prosody synthesis quality comparable to traditional methods. By enhancing the robustness of reference-based transfer tasks and improving feature disentanglement, our method paves the way for more controllable and expressive speech synthesis systems.

Research Ethics Approval

This project obtained approval from the PPLS Research Ethics Committee.
Ethics application number: 396-2324/1
Date when approval was obtained: 2024-07-16
The participants' information sheet and a consent form are included in the appendix.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Lin.Liu*)

Acknowledgements

I am deeply grateful to my supervisor, Professor Simon King, for his invaluable guidance, insightful feedback, and unwavering support throughout my dissertation. His expertise and patience were pivotal in overcoming the challenges I faced, and his encouragement played a crucial role in the successful completion of this work. I also extend my heartfelt thanks to the University of Edinburgh for fostering a supportive and enriching academic environment that greatly facilitated my research journey.

Table of Contents

Chapter 1

Introduction

This project focuses on the Prosody Transfer (PT) task, which aims to synthesize speech that mirrors the prosodic features of reference speech while preserving the content and speaker identity of the target speech. The goal is to produce expressive, natural-sounding speech. However, current models often struggle with significant performance disparities between training and inference, as well as issues like speaker leakage [1]. These challenges stem primarily from the teacher-forcing training strategy, which relies on target speech as references during training to speed up convergence. This approach, however, can lead to performance degradation during inference, especially when non-target reference speech is used, resulting in synthesized speech that is low in quality, and unclear in articulation. Additionally, the entanglement of prosody with timbre and content is exacerbated by the teacher-forcing strategy, leading to unwanted timbre alterations and content leakage.

To address these challenges, our project develops a system that employs prosodically-aligned references—carefully designed to be content- and speaker-independent yet rich in prosodic information—during training. By shifting away from the traditional teacher-forcing approach and utilizing non-target references, this strategy bridges the gap between training and inference, resulting in more consistent performance even when the reference differs from the target. A significant challenge in this field is identifying prosodically similar speech for each training sample. Misaligned references can lead to mismatches between reference and target prosody, hindering effective model training and convergence. To overcome this, we leverage unit selection techniques [2] to generate prosody-aligned references by concatenating segments that closely match the target speech's prosody. Additionally, we apply speaker normalization to speaker-relevant features to avoid considering timbre in the unit selection process. This ensures that

the model learns speaker-independent prosodic features, effectively mitigating speaker leakage and enhancing the model's ability to transfer prosody without altering the target speaker's identity.

Our study tests several hypotheses: First, if prosody is truly transferable, using prosodically-aligned references should enable the model to synthesize speech that accurately mirrors the reference speech's prosody. Second, employing non-target references during training is expected to reduce the performance gap between training and inference, particularly in pronunciation accuracy and audio quality, in contrast to the teacher-forcing strategy. Third, models trained with content- and speaker-different reference speech should learn a speaker- and content-independent prosody representation, resulting in synthesized speech that more closely resembles the target speaker compared to outputs from teacher-forcing methods.

The experiments will demonstrate that our method significantly reduces speaker leakage and delivers consistent performance, regardless of whether the reference matches the target in speaker or content. Moreover, it maintains prosody synthesis quality comparable to models trained with target references.

In the forthcoming report, the "Background" section will explore the concept of prosody, the challenges of prosody transfer, and the unit selection method we plan to use. The subsequent sections will detail the entire process and design choices involved in building the system—from selecting and preprocessing the dataset, to generating prosody-aligned references through carefully designed matching criteria, and finally training the prosody transfer model using these references. Lastly, the "Experiments" section will present the experiments conducted and discuss the results.

Chapter 2

Background

2.1 Prosody

According to [3], prosody encompasses variations in speech signals that extend beyond phonetic details, speaker identity, and channel influences. It enhances the comprehension of spoken language through the delivery of speech [4] and plays a crucial role in conveying meanings that surpass mere words, including emotions and emphasis [5]. Traditional Text-to-Speech (TTS) methods, like Tacotron2[6] and FastSpeech2[7], which receive text inputs (combined with speaker ID in multi-speaker settings), can only produce speech with averaged prosody, which reduces the naturalness of synthesized speech compared to real speech. This limitation arises because these models cannot handle the one-to-many mapping problem between text and speech, where the same text can be spoken in various ways with different intentions, leading to diverse speech outputs. Currently, reference-based prosody transfer like [3] and text-prompt guided models such as [8] are proposed to address the one-to-many mapping problem and generate speaking styles.

2.2 Prosody Transfer (PT)

Prosody Transfer (PT), introduced by [3], employs reference speech as a prosody prompt to guide TTS models in synthesizing prosody. PT models typically employ a fixed-length style embedding extracted from reference speech by a reference encoder, which is then indirectly updated through spectrogram reconstruction loss, to guide the prosody synthesis. A critical and challenging aspect of this task is to identify an appropriate prosodically-informative reference. During training, the teacher-forcing

strategy utilizes target speech as references, which perfectly aligns with the desired prosody, to guide the target speech. While effective for achieving rapid convergence and high performance in training, this method becomes less effective when the reference shifts to a different speaker or content setting, often leading to degraded performance and speaker leakage issues[1][9][3]. This issue arises primarily because the teacher-forcing strategy tends to leak ground truth information during training, prompting the model to replicate rather than truly transfer relevant prosodic features to the target speaker and text. Moreover, training to transfer prosody from ground truth speech not only transfers prosodic features but also unintentionally entangles speaker and content details, which negatively impacts the timbre and content accuracy of the synthesized speech. To address these issues, this project proposes using non-target reference speech that contains only relevant prosodic information. This setting is expected to reduce the performance mismatch between training and inference and force the model to transfer prosodic information, thereby mitigating the speaker leakage problem.

[10] highlighted a similar challenge where models trained with same-speaker, same-text settings struggle when tested with different-speaker, different-text references. They proposed training with a prosodically-similar reference, either matching in text or fundamental frequency (F0) patterns[10]. However, this approach yielded poorer outcomes compared to teacher-forcing methods, leading to the conclusion that prosody may not be effectively transferable. Despite these efforts, consistently finding a prosodically-informative utterance remains difficult, as even speech with the same text or F0 can display varied prosody. This often results in a significant gap between reference and target prosodies, pushing models to learn unachievable prosody traits and struggle with convergence. To address these issues, we advance our method by using concatenation-based techniques to generate prosody-matched speech that isn't available in the dataset, thus reducing the prosodic gap without relying on possibly nonexistent reference speech.

2.3 Unit Selection

The unit selection technique, outlined by [2], generates speech by selecting and concatenating pre-recorded segments like diphones. This method optimizes the selection by minimizing linguistic and acoustic distances to the target and reducing join costs at concatenation boundaries. With a dataset rich in phonetic and prosodic data, this approach can generate natural, human-like speech that is not originally present in the dataset, featuring varied prosody. Inspired by this concatenation-based method, this

project aims to generate prosodically-aligned speech by matching and concatenating segments that exhibit similar prosodic characteristics to the target. Compared to parametric or deep learning-based methods, unit selection offers significant advantages for this task. Firstly, it allows for great control over the speech output because the target loss can be tailored to match specific characteristics, such as F0 patterns. Moreover, this method primarily involves retrieving and concatenating speech segments, which enhances efficiency in terms of both runtime and computational costs. Thus, using unit selection ensures that creating prosody-matched speech remains a straightforward process without demanding extensive hardware resources.

Chapter 3

Dataset Selection and Preprocessing

3.1 Dataset Selection

For the prosody transfer task, it's essential to utilize a dataset that contains diverse prosody and multiple speakers, with the latter being critical for testing the transfer of speaker-irrelevant prosody. Additionally, an expressive TTS model generally requires substantial data to achieve good generalization. After excluding datasets with monotonous prosody like VCTK[11] and LJSpeech[12], and those with limited size or prosody diversity such as SAVEE[13] and RAVDESS[14], IEMOCAP[15] and ESD[16] emerged as suitable candidates due to their inclusion of emotional speech with varied prosody and multiple speakers. Despite its rich expressiveness and inclusion of both speech and non-speech sounds like laughter and silence, IEMOCAP often contains background noises such as human chatter or overlapping speech, making it unsuitable for TTS tasks[17]. Preliminary testing also confirmed that synthesized speech from IEMOCAP was too noisy for TTS applications. Consequently, we selected the ESD dataset, which is permitted for research use. It consists of 10 speakers with 350 utterances each [16]. For our experiments, we specifically focused on the English subset, excluding the Chinese subset.

3.2 Data Preprocessing

As our primary focus is on the FastSpeech2 model [7], which will be discussed in detail in later chapters, it predicts prosody based on pitch, energy, and duration. Therefore, we preprocess the dataset according to FastSpeech2's pipeline. This process begins with using Short-Time Fourier Transform (STFT) to extract the mel spectrogram from

the waveform. Pitch is then extracted using the pyworld library, energy is computed by summing the squared magnitudes of the spectrogram over time, and phone- and word-level durations are obtained using the Montreal Forced Aligner (MFA)[18].

Chapter 4

References Generation

The concatenation-based prosodically-aligned references generation process involves three primary aspects: selecting the appropriate unit type, extracting prosodically-related features, and designing matching criteria.

4.1 Unit Type Selection

In determining the appropriate unit type for segmentation and concatenation, we opted for word-level units, as prosody is generally observed at higher levels of representation, such as syllables, words, or utterances. We used the Montreal Forced Aligner (MFA) to align speech with word-level transcriptions, then segmented the waveform accordingly to extract the units. Unlike traditional unit selection methods that often rely on diphone-level features to ensure smooth transitions [19], our approach does not require this level of granularity, as the generated speech is solely intended to provide prosody prompts and does not need to sound natural.

4.2 Prosody Feature Extraction

The second step involves extracting prosodic features to quantify the prosody distance between units. We explored two methods in our experiments: the acoustic-based method, which includes pitch, energy, and duration, and the spec-based method, which leverages low-frequency spectrogram bins. While the spec-based method offers a broader spectrum of prosodic information, the acoustic-based method is more computationally efficient and provides stable results for prosody matching.

4.2.1 Acoustic-Based Approach

In the acoustic-based approach, we focus on three primary prosodic features: pitch, energy, and duration[20]. Pitch captures tonal variations that help distinguish between questions and statements, with questions typically ending in a higher F0 compared to statements [21]. Energy reflects emphasis and stress[22], highlighting key words or phrases in speech, while duration measures the length of phonemes and pauses, playing a crucial role in the perception of rhythm and flow[23]. Together, these features provide a comprehensive view of prosodic elements, enabling a deeper analysis of speech characteristics. During our dataset preprocessing, we extracted these features by computing and averaging pitch and energy within word boundaries to obtain word-level values.

4.2.1.1 Why three types of acoustic features?

Our model was initially built using only F0, given its close alignment with human auditory perception and its importance in capturing the pitch and intonation of speech[24]. However, this approach led to suboptimal prosody matching between the target speech and the generated prosodically-aligned speech, creating several issues.

Firstly, the lack of duration constraints resulted in mismatches in timing, such as aligning the word ”arrows” with ”i,” or omitting essential pauses, leading to clearly prosodically dissimilar pairings. This issue stemmed from obtaining word-level features through averaging, which inadvertently removed crucial duration information. A similar problem also existed in the spec-based methods, necessitating the application of similar duration constraints. In addition, concatenated neighboring units often exhibited mismatched amplitudes, leading to disjointed sounds and disrupting prosody continuity.

To address these challenges, we introduced duration and energy as additional features for prosody matching. Specifically, word-level durations and word-level mean energy values were categorized into 10 evenly distributed classes. Only word segments within the same duration and energy class were allowed to match. The choice of 10 classes was intentional, aiming to balance precise duration and energy matching with ensuring an adequate number of units in each class for effective selection. Moreover, pauses in the target speech were preserved to maintain consistency. Personal listening tests confirmed that these constraints significantly reduced mismatched segments and led to smoother, more harmonious prosody without abrupt transitions.

4.2.2 Spec-Based Approach

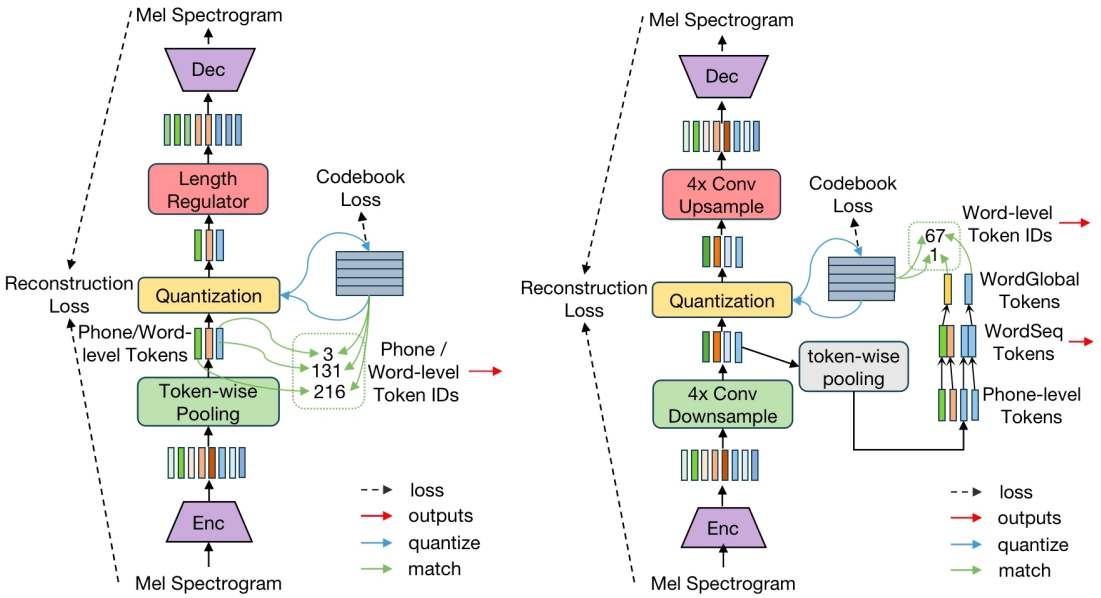
For the spec-based methods, inspired by [25] and [26], we focused on the low-frequency portion of the spectrogram to isolate prosody-related features while filtering out most of the speaker and content information. This was confirmed by applying a low-pass filter, which rendered the speech nearly unintelligible regarding content and speaker identity but preserved the prosodic characteristics, akin to speaking with a muffled mouth. In our experiment, we targeted frequencies below 400Hz. To effectively encode these prosodic features, we employed a vector quantization variational autoencoder (VQ-VAE) [27], converting spectrograms into word-level tokens that capture essential prosodic elements while reducing dimensionality through reconstruction and quantization losses.

4.2.2.1 Building the VQVAE structure.

The VQVAE model encodes spectrogram features into frame-level embeddings that correspond to discrete tokens. These sequential embeddings are then averaged into word-level values based on detected word boundaries, with each value replaced by the closest token index from the codebook. By employing the VQ-VAE framework, this approach facilitates the encoding of spectrograms into discrete features, making it practical to measure similarity between embeddings and efficiently compute distances between tokens.

The VQ-VAE structure typically consists of three main components: an encoder, a vector quantization (VQ) module, and a decoder. To obtain word-level features, there are two primary approaches: training the VQ module directly at the word level or training at finer-grained levels (such as phoneme or frame levels) and then pooling these to form word-level representations. Inspired by [26], we initially configured the VQ-VAE with token-wise pooling and a length regulator positioned before and after the VQ module, respectively, as shown in Fig4.1a. This design enabled the conversion of frame-level representations into word or phoneme levels based on duration boundaries.

However, our experiments revealed several challenges. Training the VQ-VAE directly at the word level required extensive pooling from frame-level to word-level representations within the encoder, complicating the reconstruction process and causing a mode collapse in the VQ module, where only a single code would activate. Although training at the phoneme level avoided this mode collapse, it often resulted in inaccurate re-synthesized spectrograms due to similar averaging issues, indicating that the model failed to capture accurate prosodic patterns. Training at the frame level proved effective



(a) VQ-VAE with token-wise pooling and length regulator (b) VQ-VAE with 4x downsampling and upsampling.

Figure 4.1: Comparison of different VQ-VAE architectures. Left: VQ-VAE with token-wise pooling and length regulator. Right: VQ-VAE with 4x downsampling and upsampling.

for spectrogram reconstruction but led to significant information loss when these features were averaged into word-level representations during inference. Additionally, this method was inadequate because it encoded individual frames without capturing the necessary contextual information, which is crucial for accurately representing prosodic patterns.

To overcome these issues, we replaced the token-wise pooling and length regulator with 4x convolutional downsampling and upsampling layers, as illustrated in Fig4.1b. During inference, quantized embeddings were generated every four frames and then averaged according to phoneme and word boundaries to derive word-level token IDs, thereby reducing the extent of pooling required. This revised approach strikes a balance between capturing contextual information and maintaining detail, resulting in a more accurate prosody representation at the word level.

4.2.2.2 What do low frequency bins encode?

We evaluate the VQ-VAE's encoding of low-frequency bins to ensure it primarily captures prosodic features without embedding unwanted speaker identity or text information. To assess this, we use hexbin plots to visualize the distribution of VQ tokens

concerning both speakers and phonemes, as illustrated in Figure 4.2. The shading within each hexagon indicates the frequency of occurrences, allowing us to observe the relationship between these tokens and speakers or phonemes.

(a) Hexbin Graph for VQ Tokens and Speakers.

(b) Hexbin Graph for VQ Tokens and Phoneme Types.

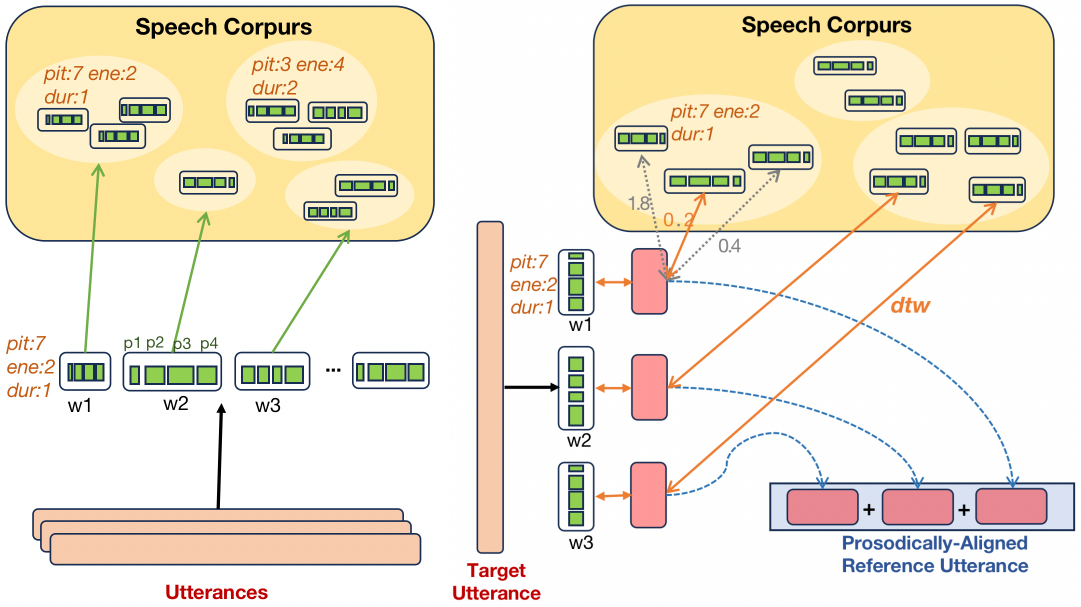
Figure 4.2: Hexbin graphs showing the relationships between VQ tokens and various linguistic features. Subfigures illustrate the density of data points for VQ tokens with speakers (a), and phoneme types (b), with higher color intensity indicating greater data concentration.

In Figure 4.2a, the evenly distributed horizontal stripes indicate that VQ tokens are uniformly spread across different speakers, with darker regions signifying higher usage frequency of specific tokens. This uniform distribution suggests that the tokens are not tied to any particular speaker. Similarly, Figure 4.2b shows that, while some darker areas imply minor encoding linked to certain phonemes, the overall token distribution remains consistent across various phonemes. This confirms that the VQ model effectively encodes features that are independent of speaker identity and content, aligning with our objective to focus on prosodic characteristics.

4.3 Matching Criteria Establishment

Establishing effective matching criteria involves developing a precise method for measuring distances between prosodic units and selecting the best units for concatenation based on minimal distance. As demonstrated in Figure4.3, in our approach, we begin by categorizing features into word-level discrete classes, using either quantized token IDs or acoustic bins (such as pitch, duration, and energy). Unit selection is restricted

within these classes to maintain consistency. Within each class, Dynamic Time Warping (DTW) is applied at the phoneme level between paired word segments, allowing for the stretching and aligning of phoneme sequences to accurately compute distances. The word segment with the lowest total distance to the target word is then identified as the best match. These matched word segments are concatenated directly without the joint smoothing typically used in traditional unit selection, as the goal here is to provide prosody-related information rather than to produce a naturally smooth reference.



(a) Constructing a prosodic corpus using word segments.

(b) Matching prosody to align with the target.

Figure 4.3: Generating prosodically-aligned references to the target. In the figures, ”pit”, ”ene”, and ”dur” are abbreviations for ”pitch”, ”energy”, and ”duration”, respectively. ”w1, w2, w3,...” and ”p1, p2, p3,...” represent words and phonemes.

The following sections will provide a detailed exploration of the methods and strategies we employed to enhance the efficiency and accuracy of DTW-based word matching. All evaluations are based on personal listening tests. Each experiment builds sequentially on the previous one, incrementally refining the algorithm. For preliminary evaluations of prosody transfer using prosodically-aligned references, we utilized the traditional FastSpeech2 architecture enhanced with Global Style Tokens (GST), as proposed by [3]. This setup establishes a baseline for assessing the effectiveness of the prosodically-aligned references, with the detailed architecture and final design choices for prosody transfer to be thoroughly discussed in the subsequent section.

4.3.1 Direct vs. Component-Based Matching.

To compute prosodic distances between word-level units, two primary approaches can be used: direct comparison of word-level features or a more granular breakdown into phoneme-level features. In the direct comparison approach, word-level units are categorized into classes based on acoustic bins or token IDs, with matching segments randomly selected within each class. The component-based approach refines this process by first grouping word segments into these classes to constrain matching, then further dividing them into their phoneme components for finer comparison. Dynamic Time Warping (DTW) is employed to align these phoneme components and calculate distances based on acoustic features or quantized embeddings. The segments are then ranked by their similarity to the target, with the closest matches chosen for use.

Experimental results indicate that relying solely on broad classifications of word-level representations and randomly matching segments within these classes is insufficient for achieving accurate prosodic alignment, often resulting in references that significantly deviates from the target prosody. Additionally, word-level features frequently fail to capture finer prosodic nuances, such as the subtle rises in intonation within question words. In contrast, the component-based method offers substantial improvement by matching more precisely and effectively capturing these subtle variations, resulting in references with more consistent prosodic patterns.

4.3.2 Matching Priorities

The selection criteria prioritize finding the closest match that differs from the original segment, belongs to a different speaker, and hasn't been overused. This approach prevents the algorithm from defaulting to identical segments due to zero distance or favoring segments from the target speaker due to similar prosody. By doing so, it ensures the use of non-target references and facilitates the transfer of speaker-independent prosody.

In cases where no suitable match meets these criteria, unmatched candidates are added to a fallback list. If necessary, the algorithm selects the first candidate from this list, which may occasionally result in matching segments from the target speaker or segment. However, this occurs in less than 3% of cases, making it an acceptable compromise for maintaining the overall objective of transferring speaker-independent prosody. This method greedily matches segments, supports sparse distribution of segments, and ensures that every word segment has at least one match, even if it

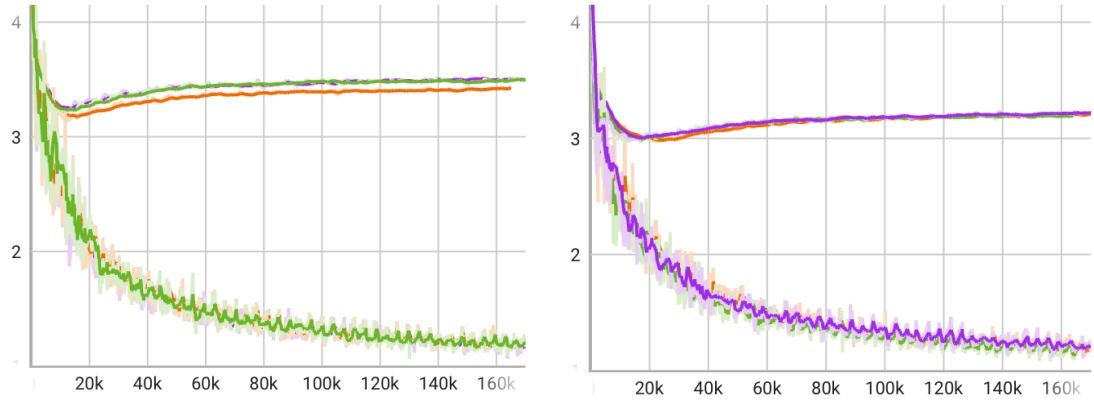
occasionally defaults to a self-match in rare instances.

4.3.3 Single vs. Multiple Speaker Concatenation

To select and concatenate speech segments for a target utterance, segments can be sourced from the target speaker, a random non-target speaker, or multiple speakers. In the random single speaker scenario, the process involves calculating a prosodically-aligned utterance for each non-target speaker and selecting the speaker whose segments have the lowest overall distance to the target as the final match. Alternatively, when using segments from the target or multiple speakers, the matching process can either restrict the selection to segments from the target speaker or allow segments to be chosen from any speaker without restrictions.

Using segments from the target speaker has the benefit of producing speech with consistent prosody, as the same speaker generally maintains uniform speaking habits. This approach also significantly reduces computational costs and runtime by avoiding cross-speaker comparisons and limiting the number of candidates per class, which minimizes the need for complex DTW computations. However, this method risks leaking speaker-specific information, which goes against the goal of focusing solely on prosody for effective feature disentanglement. Despite these concerns, the same-speaker setting serves as a useful baseline to assess whether the model can effectively transfer prosody across different speakers while minimizing speaker leakage. Compared to the multi-speaker setting, the single-speaker setting produces reference with a more cohesive style, avoiding the abrupt transitions that can occur due to timbre differences across speakers. Nevertheless, despite its challenges, the multi-speaker setting encourages the model to ignore speaker identity and focus exclusively on transferring speaker-independent prosodic features.

When comparing the performance of prosody transfer using prosodically-aligned references across different speaker settings, fig 4.4 shows that all settings had similar total loss during training. Personal listening tests also suggest that all three settings delivered comparable prosody transfer. However, the same-speaker setting often resulted in unclear and low-quality speech, especially when there was a significant timbre mismatch between the target and reference speakers. In contrast, the single-speaker and multi-speaker settings were more robust to timbre variations, with the multi-speaker setting offering slightly better timbre preservation. Despite this advantage, the multi-speaker setting demands significantly larger DTW matrix computations within the



(a) Total Loss for spec-based method, across different speaker settings.

(b) Total Loss for acoustic-based method, across different speaker settings.

Figure 4.4: Comparison of total loss across different speaker settings (single, same, multi) for (a) acoustic-based and (b) spec-based methods. The legend indicates the speaker settings used in both subfigures.

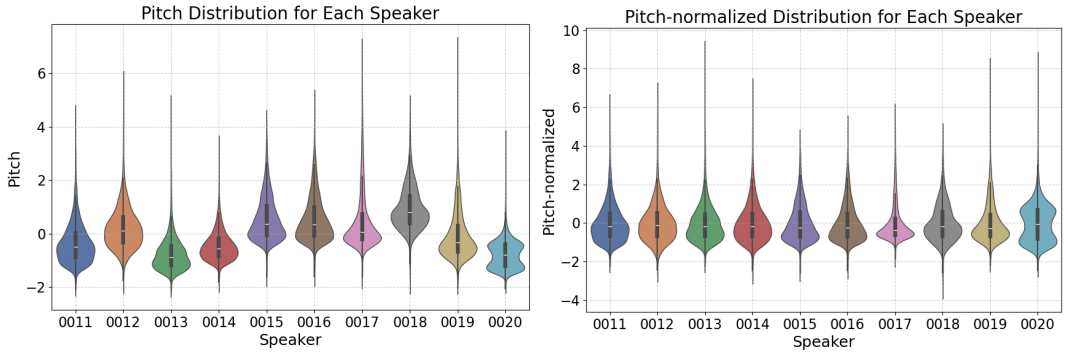
same class due to the increased number of possible segment matches across multiple speakers, which exponentially increases the complexity and processing time. To balance computational cost and performance, the single-speaker setting is chosen for further experiments.

4.3.4 Speaker Normalization

When testing prosody transfer performance using a prosodically-aligned reference, the spec-based method shows better speaker preservation. This is because the quantized tokens effectively exclude speaker-relevant features, as discussed in Section 4.2.3. Additionally, the matching process ensures the reference speaker is different from the target speaker, further reducing speaker leakage. In contrast, the acoustic-based method exhibits a level of speaker leakage similar to that of the model trained with ground truth references. This finding challenges our initial assumption that using a non-target speaker's reference speech would naturally reduce speaker leakage in prosody transfer.

One potential explanation for this discrepancy is that f0, being closely tied to the unique physical characteristics of a speaker's vocal folds, naturally carries speaker-specific information. To test this hypothesis, we visualized the word-level pitch values for each speaker, as shown in Figure 4.5a. The visualization clearly shows that different

speakers have distinct f0 distributions. Thus, although the matching process was designed to avoid same-speaker pairing, matching based on the lowest DTW value across features that include f0 often results in speakers with similar timbres being paired together. This unintended consequence leads to the leakage of speaker information during training. This issue is further corroborated by Figure 4.6a, where the uneven color distribution highlights frequent matches between two speakers with similar pitch profiles, indicating that the matching process is heavily influenced by timbre similarities.

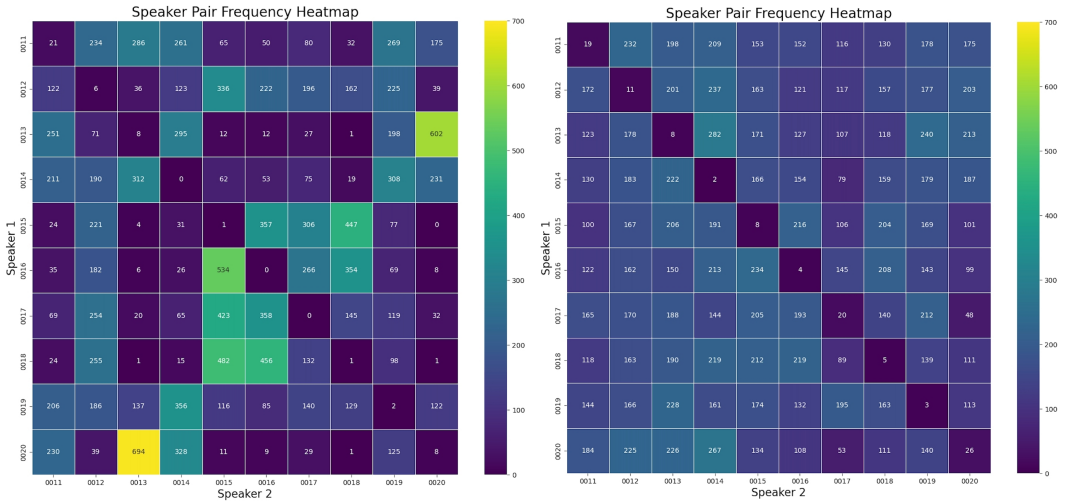


(a) Pitch distribution per speaker (Violin Plot). (b) Normalized Pitch distribution per speaker (Violin Plot).

Figure 4.5: Comparison of pitch distributions per speaker before and after normalization. Both plots reveal the central tendency and variability through points (medians) and rectangles (interquartile ranges).

To address this issue, we perform speaker normalization on speaker-related features, specifically F0 (fundamental frequency), to standardize pitch variations across different speakers and minimize inter-speaker variability. Specifically, we begin by normalizing the speaker's phone-level f0 by subtracting the speaker-specific mean and dividing by the speaker-specific variance. We then apply min-max normalization to all features, including the normalized pitch, scaling them between 0 and 1 to ensure each feature contributes equally in the DTW computation. Following this, we perform token-wise pooling on the normalized phone-level values based on word boundaries, creating word-level values that are then binned for matching constraints. This process ensures that both the matching constraints and DTW calculations are applied to features with speaker-specific information minimized, effectively reducing timbre differences.

As shown in Figure 4.5b, the normalization process leads to a more uniform pitch distribution across speakers. The corresponding heatmap in Figure 4.6b exhibits a more uniform color distribution without extreme variations, indicating a significant reduction



(a) Speaker pair frequency before speaker normalization.

(b) Speaker pair frequency after speaker normalization.

Figure 4.6: Comparison of matching frequencies before and after f0 normalization. The left heatmap shows the raw frequencies of speaker1's target speech matched to speaker2's reference speech, while the right heatmap shows the normalized frequencies. Color intensity and annotated values indicate the interaction frequencies between speakers.

in speaker-specific information during matching. Furthermore, the prosody transfer performance after applying speaker normalization demonstrates improved speaker preservation compared to methods without normalization, highlighting the effectiveness of this approach in mitigating speaker-specific information leakage.

4.3.5 Efficiency Control

The Dynamic Time Warping (DTW) process, which calculates distances between segment pairs across entire datasets, is computationally intensive due to the large matrices involved. To reduce this burden, we implement several preprocessing steps to optimize the process.

First, high dimensionality poses a significant challenge to processing speed. The acoustic-based method benefits from using lower-dimensional representations, where each phoneme-level unit is represented by a single value for each acoustic attribute. In contrast, the spec-based method typically employs 128-dimensional embeddings for the phonemes that make up words, leading to highly detailed representations. To manage this complexity, we reduce the dimensionality by a factor of eight through pooling,

making the processing more efficient.

Next, reducing the size of DTW matrices is critical for optimizing computational efficiency. To achieve this, we prioritize single-speaker matching over multi-speaker settings, where cross-speaker comparisons are avoided. Besides, instead of applying DTW to all paired word representations, we first categorize these representations into broad groups based on word-level acoustic bins or quantized token IDs. DTW is then applied only within these specific groups, resulting in fewer candidates per class and thus smaller computational matrices, reducing the overall computational load.

However, balancing efficiency and matching performance is crucial, as reducing the number of candidates per class can limit options and increase the risk of same-speaker or self-matching, which is undesirable. In acoustic-based method, we prefer equal-frequency binning over equal-width binning to achieve more uniform divisions. This approach ensures that data is evenly distributed across bins, reducing the likelihood of creating overly large matrices that waste computational resources or excessively small matrices that lack sufficient candidates for effective matching. After implementing these strategies, if the matrix size still exceeds 400x400, we will process the data in batches to further optimize computational efficiency.

4.4 Acoustic-Based vs. Spec-Based References

After establishing the overall matching criteria, we compared the prosodically-aligned references generated by acoustic-based and spec-based methods. All evaluations were conducted through personal listening tests.

Spec-based method often faces challenges in producing speech that closely matches the prosody of the target. It frequently results in prosody mismatches, such as pairing question intonations with statement intonations. In contrast, acoustic-based methods demonstrate greater robustness, consistently producing references with prosody that closely aligns with the target, with differences primarily in timbre. This discrepancy in performance is likely due to the unstable and less transparent nature of low-frequency encoding in spec-based methods, compared to the more reliable signal processing techniques used to extract acoustic features. Furthermore, spec-based methods are more complex in terms of matching efficiency, requiring VQVAE training and the encoding of low-frequency bins, which generates high-dimensional data. This, in turn, increases the runtime and computational resources needed for DTW computation.

Hence, the comparison indicates that the acoustic-based method outperforms the

spec-based method in both prosody similarity and computational efficiency. As a result, the acoustic-based approach will be adopted for our subsequent experiments.

Chapter 5

Prosody Enhanced TTS model

In this section, we enhance the FastSpeech2[7] model to transfer prosody by incorporating reference speech, resulting in a prosody-enhanced TTS model. We evaluate different model architectures, including FastSpeech2 with Global Style Tokens (GST)[3], Local Style Tokens (LST), and their combination, to improve prosody modeling. The results show that combining FastSpeech2 with both GST and LST enhances prosody modeling, leading to more expressive speech, though caution is required to prevent overfitting.

5.1 FastSpeech2

In this project, we utilize the FastSpeech2 model, as described by [7], as our TTS architecture. FastSpeech2 is engineered to predict prosody by analyzing duration, pitch, and energy based on phoneme inputs. However, the model's limitation in handling the one-to-many mapping between text and spoken speech often results in an averaged prosody reflective of the training dataset, thus hindering the expressiveness and naturalness of the synthesized speech. To address this, we introduce a reference speech and employ an additional encoder to integrate the reference, thereby guiding the prosody synthesis.

5.2 GST and LST

Preliminary experiments on prosody transfer using prosodically-aligned references suggest that while GST-based TTS can generally replicate the prosody of the reference speech, it often produces a uniform prosody across utterances. This uniformity results in a mechanical tone, lacking in detailed nuances such as distinct emphasis, pauses, and

varied speaking rates, which are essential for expressive, natural-sounding speech. In contrast, the acoustic-based concatenated reference more accurately captures the target speech's prosody, including its finer details. This indicates that the limitations may not stem from the reference speech's ability to convey prosodic cues, but rather from the GST model itself. The GST model uses a fixed-size global embedding from the reference speech, applying it uniformly across the speech, which often leads to a more generalized and less nuanced prosody representation.

Thus, to enhance the detail of prosody in synthesized speech, we explore a structure similar to Global Style Tokens (GST) called Local Style Tokens (LST), as illustrated in Fig5.1. Both GST and LST utilize style embeddings extracted from reference speech to guide the synthesis process through a weighted combination of learnable tokens. However, unlike GST, which captures broad, utterance-level features, LST focuses on fine-grained prosodic nuances at the intra-utterance level. By capturing these subtle variations, LST enables the reproduction of more expressive and natural-sounding speech. These fine-grained prosodic embeddings are then aligned with phoneme representations using attention mechanisms and integrated into the FastSpeech2 pipeline. While our LST structure shares similarities with the method described in [28], it differs in a key aspect: we deliberately avoid incorporating text information into the prosody space construction, ensuring that the prosody space remains purely prosodic until it is aligned with phonemes.

5.3 Comparison of Prosody Enhancement Techniques

Experiments were conducted to evaluate four versions of the model architecture: the baseline FastSpeech2, FastSpeech2 augmented with either GST or LST, and FastSpeech2 enhanced with both GST and LST modules. The reference speech at training and validation stage both utilizes acoustic-based references concatenated with single speaker setting, and all other settings remain unchanged to ensure consistent experimental conditions across various tests. Evaluations are based on training graph analysis and personal listening tests.

The training graph in Fig5.2 shows that FastSpeech2 exhibits significantly higher training and validation losses compared to the other configurations. It also demonstrates considerable fluctuations, particularly in pitch and duration losses, indicating the model's difficulty in capturing variations in these features. The other three methods, which incorporate additional encoders for reference speech, display similar validation

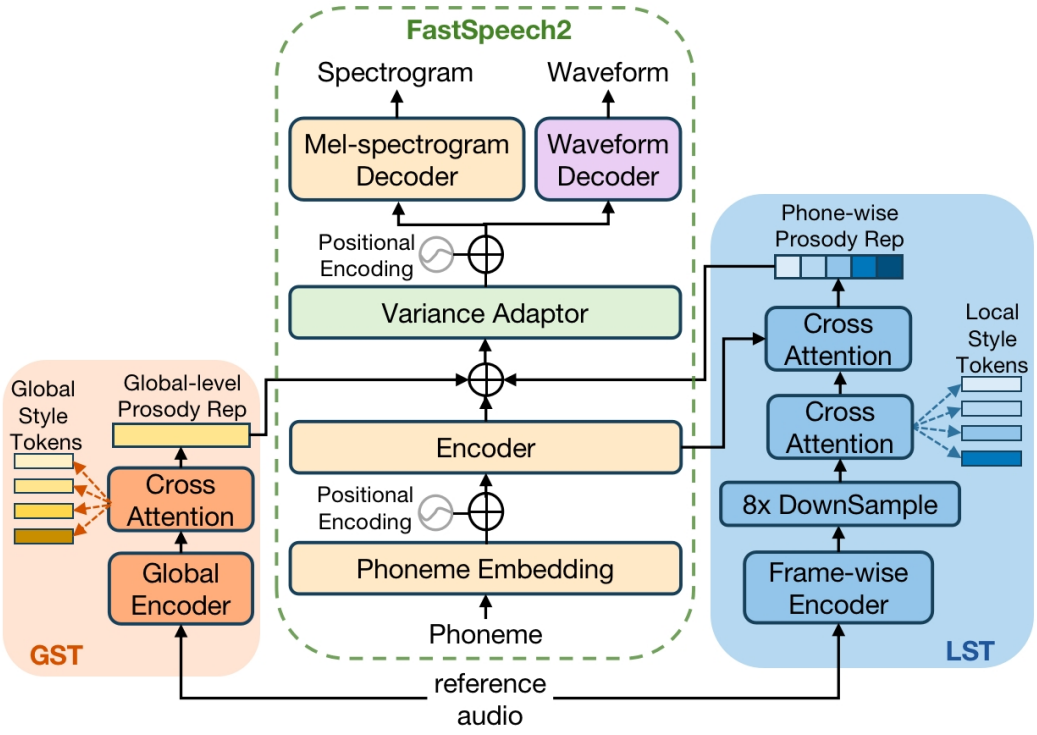
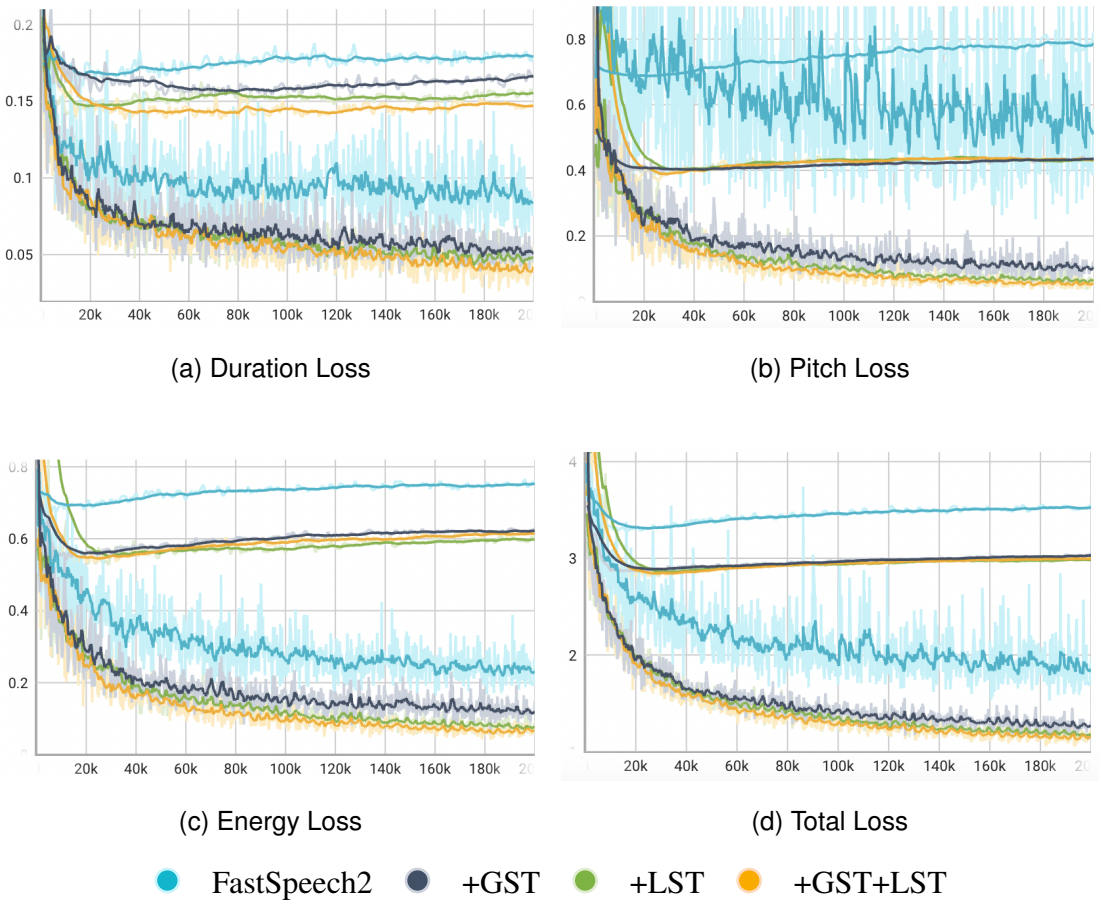


Figure 5.1: Prosody Enhanced TTS model structure: FastSpeech2 with GST and LST

performance. Notably, LST, whether used alone or in combination with GST, shows faster training and a better fit to the training data, with significantly lower training loss.

Based on my listening assessments, the FastSpeech2 model produces a mechanical tone with a consistent speaking rate and uniform tones across words. The GST-only model demonstrates stable performance and robustness across various scenarios but tends to exhibit averaged prosody, lacking expressiveness and the ability to capture nuanced prosodic variations. The LST-based model, though slightly better at prosody, often suffers from unnatural and unclear pronunciation, making it the least favorable among the four versions. In contrast, the combination of GST and LST shows more variation within utterances and improved prosody expression, suggesting that LST can effectively supplement GST by adding finer details. However, LST's performance is not as robust as GST's, and caution is needed when using LST, as models equipped with it are prone to overfitting. For example, when trained with non-target references lacking sufficient prosodic similarity or with ground truth references containing extraneous features, LST can overfit to irrelevant details, leading to unnatural prosody during validation despite better fitting to the training data. In our model configuration, the use of the LST mechanism is appropriate since the reference guides only prosody-related features, thus avoiding the fitting of unwanted characteristics.



(a) Duration Loss

(b) Pitch Loss

(c) Energy Loss

(d) Total Loss

Figure 5.2: The training graph for FastSpeech2, combined with GST or LST, as well as GST and LST, shows total loss, pitch loss, energy loss, and duration loss. The bottom and upper lines of the same color represent training and validation performance, respectively.

Considering the experimental results, the combination of FastSpeech2 with GST and LST proves more favorable for this task and will be utilized for further evaluation.

Chapter 6

Ground Truth vs. Prosodically-Aligned References in Prosody Transfer

At this stage, we have fully established the system for generating prosodically-aligned references and configuring the prosody transfer model. We now evaluate the model trained with our prosodically-aligned references (prosodic-based) against the one trained with ground truth references (gt-based), focusing on hypotheses related to the training-validation performance gap, prosody transfer accuracy, and timbre preservation.

We evaluate the prosody transfer model under three training configurations: using ground truth references (gtPt), prosodic references (prosodicPt), and random references (shufflePt). The shuffle-based model, which disrupts prosodic alignment by using random references during training, serves as a baseline. For inference, we use two types of references: ground truth (gtRef) and randomly selected non-target references (shuffleRef). To avoid extreme mismatches in prosody, the shuffle references are constrained to ensure that the text length difference between the reference and the synthesized output falls within a 2:1 or 1:2 ratio. All references and texts used for evaluation are unseen during training or validation and are drawn from the test set. We explore various model-reference combinations, such as 'gtPt-shuffleRef,' where a model trained with ground truth references is inferred using random different references. This naming convention will clarify the different setups used in the subsequent experiments.

The evaluation process includes multiple approaches, incorporating objective metrics like Word Error Rate (WER) and speaker similarity (SIM), as well as subjective evaluations through both personal and formal listening tests. For objective evaluations, we follow the pipeline established in SeedTTS [29]. In the formal listening tests, we recruited 20 native speakers from the US and UK via Prolific [30], each completing

a 30-minute listening session. The listening test is divided into three parts, each assessing a specific aspect of model performance: synthesis quality, prosody transfer, and speaker identity preservation. Speech samples were randomly selected and then manually screened to exclude those with quality issues unrelated to our approach, such as difficulties in synthesizing certain phonemes due to limited dataset coverage or incomplete reference speech caused by alignment and segmentation errors.

6.1 Performance Gap

The model trained with ground truth references tends to perform well during training, but its performance degrades during testing when the reference differs in speaker and content from the target. This degradation includes noisy output, unnatural prosody, and difficulty in clearly synthesizing words. This leads to our first hypothesis: using non-target references during training, which more closely align with test conditions, should reduce the performance gap between training and inference compared to the traditional teacher-forcing training strategy. The performance gap encompasses various aspects, including intelligibility, naturalness, and audio quality.

To validate this hypothesis, we will compare the performance gap between the gt-based model and the prosodic-based model when using ground truth versus shuffle references. We will begin by conducting my personal listening tests, followed by a more formal evaluation using Word Error Rate (WER) to assess intelligibility and expert listening tests to evaluate perceived quality.

Training Graph Analysis. As shown in Figure 6.1, which compares three models trained and validated using different reference types, the gt-based model achieves the best performance and convergence, with significantly lower loss. In contrast, the shuffle-based model exhibits substantial fluctuations and the poorest performance, likely due to overfitting to noisy references, leading to a failure to converge. The prosodic-based model, which uses references designed to capture prosodic features, strikes a balance between the two extremes, suggesting it effectively captures meaningful information that enhances learning.

Personal Listening Test. From my personal listening test, the gt-based model delivers the best performance, with high speech quality and fidelity with ground truth references. However, its performance degrades significantly when using non-target references, introducing noise and a robotic tone, particularly when the reference timbre differs distinctly from the target. Additionally, it is more prone to unclear pronunci-

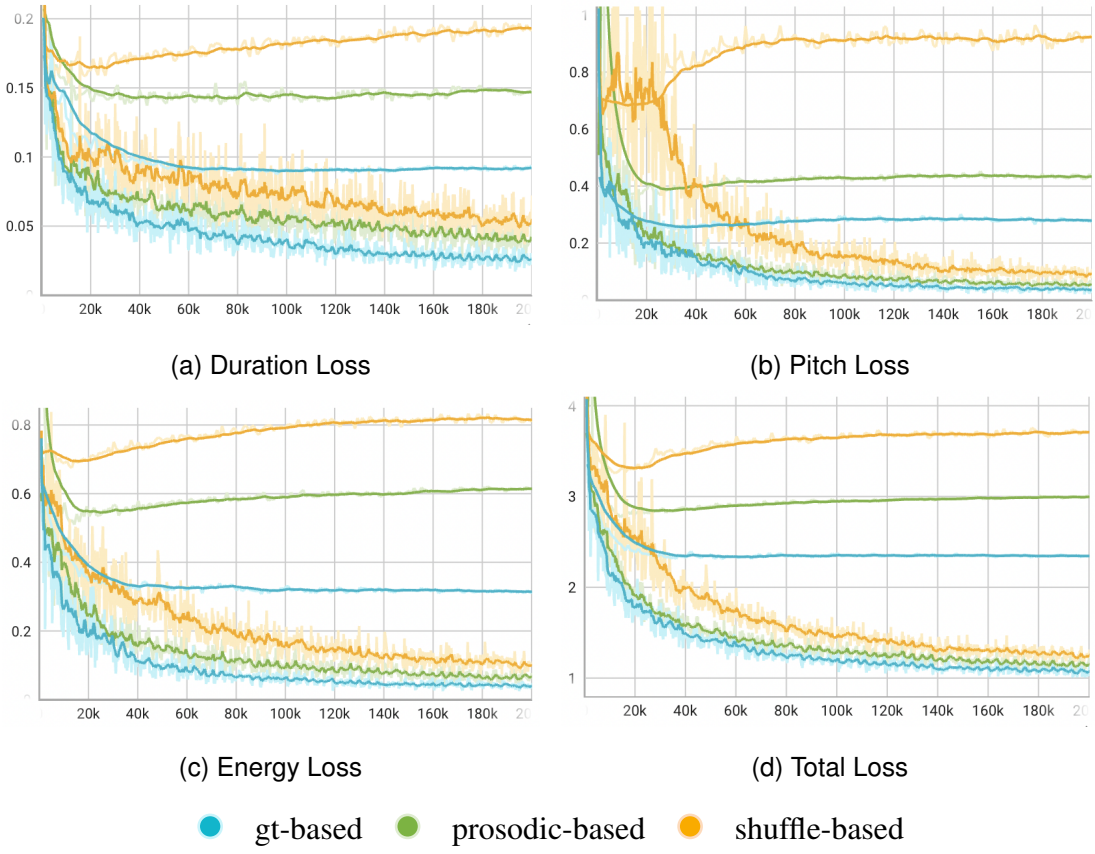


Figure 6.1: The training graph for prosody transfer compares models trained and validated with ground truth (gt-based), prosodically-aligned reference (prosodic-based), and random reference (shuffle-based). It displays total loss, pitch loss, energy loss, and duration loss, with the bottom and top lines of each color representing training and validation performance, respectively.

ations and murmurs. In contrast, the prosodic-based model trained with non-target references, while exhibiting slightly lower quality and less accurate replication of the reference compared to the gt-based model when using ground truth references, maintains comparable quality and clear pronunciations across both target and non-target settings, demonstrating much more robust performance and clarity.

Word Error Rate. To evaluate the intelligibility gap between the gt-based and prosodic-based models, we measure the Word Error Rate (WER) using the *Whisper-large-v3* model [31]. We also calculate the WER for the vocoded ground truth (gtvoc) to establish a baseline for the highest possible intelligibility. The results are presented in Table 6.1. The gt-based model shows a significant WER increase when using

non-target references compared to ground truth references, indicating a performance drop. In contrast, the prosodic-based model maintains consistent WER across different reference types, with only a slight decrease compared to the gtPt-gtRef configuration. This supports the hypothesis that using non-target references can mitigate unclear pronunciations during inference compared to applying teacher-forcing training strategy.

Model	**gtRef**	**shuffleRef**
gtPt	28.48%	35.91%
prosodicPt	29.89%	29.54%
Baseline (gtvoc)	13.37%	

Table 6.1: WER for Different Model and Reference Combinations

Listening Test. A Mean Opinion Score (MOS) test using a 5-point Likert scale was conducted to evaluate the perceptual quality of the gt-based and prosodic-based models. A vocoder-processed version of the ground truth speech was included to account for any degradation caused by the vocoder, providing a benchmark for the highest achievable MOS. We chose not to test the original ground truth, as vocoder-induced degradation was not relevant to our evaluation focus, and thus this component was excluded from the assessment. Based on the results presented in Table 6.2, we observe that the prosodic-based model achieves higher MOS scores compared to the gt-based model, whether using ground truth references (gtRef) or random references (shuffleRef). Specifically, the prosodicPt configuration scores 3.12 with gtRef and 2.92 with shuffleRef, indicating consistent performance across different reference types. In contrast, the gtPt model shows a significant drop in MOS, from 2.91 with gtRef to 2.20 with shuffleRef, underscoring its sensitivity to reference variability.

Configuration	**gtRef**	**shuffleRef**
gtPt	2.91 ± 0.51	2.20 ± 0.41
prosodicPt	3.12 ± 0.52	2.92 ± 0.55
Baseline	**gtvoc: 3.80 ± 0.49**	

Table 6.2: Mean Opinion Scores(MOS) for Different Model and Reference Configurations

The evaluation results consistently support our hypothesis that using non-target ref-

erences during training helps maintain consistent performance across different reference types in terms of intelligibility, naturalness, and audio quality.

6.2 Prosody Matching

In the prosody transfer task, the objective is to transfer speaker- and content-independent prosodic features to target speech. We hypothesize that if prosody is transferable, using prosodically-aligned references should enable the model to synthesize speech that accurately reflects the reference speech's prosody. To validate this, we compare the prosody matching performance of models trained on ground truth, prosodically-aligned references, and shuffled references, all inferred using non-target references. The shuffle-based model, trained with random references, was expected to produce the least accurate prosody and served as the baseline in this evaluation.

For this evaluation, we will primarily rely on subjective metrics, including my personal and formal listening tests. No objective evaluation of prosody matching is conducted because, in the context of prosody transfer from a different reference speech, there is no target speech available for traditional comparison methods, such as calculating pitch or periodicity error.

Personal Listening Test. In my personal listening tests, the gt-based model achieves the closest match in prosody to the reference speech. However, it tends to transfer features too broadly, mixing prosody with text and timbre. As a result, the output may sound like the speaker is imitating someone else's voice, which can lead to an unnatural or overly stylized sound that doesn't quite capture the original intent of the reference speech. On the other hand, the prosodic-based model successfully captures distinct prosodic variations, such as increasing tonal shifts. However, compared to the GT-based model, it lacks some expressiveness and can sound flat. This is likely due to the less precise prosodic matching during training compared to the gt-based model. Nonetheless, this slight compromise in prosody is acceptable as it is generally challenging to distinguish the prosodic differences between samples inferred by the GT-based and prosodic-based models. In contrast, the shuffle-based method produces synthesized speech with prosody that is distinctly different from the other two models. It often generates speech with mismatched prosody compared to the reference, making it easy to distinguish through aspects such as overall emotional tone, speech rate, and tonal variation. It highlights that prosody can be effectively transferred to the target through training with prosody-aligned references.

Listening Test. To evaluate prosody transfer performance, we conducted a MUSHRA-like test where participants compared samples synthesized by the gt-based, prosodic-based, and shuffle-based models side-by-side, all inferred using non-target references. To guide participants focus solely on prosody similarity to the reference, we excluded speech samples with very poor quality, particularly those that struggled with clear word synthesis. Participants then rated each sample on a 100-point scale, assessing how well the prosody matched the reference. The evaluation focused on key aspects of prosody, such as pitch (e.g., rising or falling patterns), rhythm (e.g., pauses and syllable timing), intonation (e.g., overall melody), and stress (e.g., emphasis on specific words). The shuffle-based model, anticipated to perform the worst, served as the anchor in this mushra-like test. The results in Table 6.3 support the hypothesis that prosody is transferable. The prosodic-based model achieves prosody similarity comparable to the gt-based model, both of which score significantly higher than the shuffle-based model. This demonstrates that our model, trained with content- and speaker-independent prosodically-informative references, can still effectively learn and replicate the prosodic characteristics of the reference speech, even without using ground truth during training.

Configuration	**Prosody Similarity**
gtPt-shuffleRef	49.14 ± 5.59
prosodicPt-shuffleRef	49.94 ± 6.54
shufflePt-shuffleRef	38.80 ± 6.85

Table 6.3: Prosody similarity scores for different model configurations.

Both listening tests validate that utilizing prosodically-aligned references instead of ground truth for prosody transfer training enables the model to effectively learn and transfer prosody, achieving performance comparable to the gt-based model.

6.3 Speaker Preservation

Speaker leakage is a common issue in prosody transfer tasks, where synthesized speech may resemble the source speaker. This likely occurs because acoustic features are inherently entangled, and the teacher-forcing training strategy further contributes to the transfer of additional features like speaker timbre alongside prosody. We hypothesize that using speech samples from speakers different from the target can help preserve the

target speaker's timbre during prosody transfer. However, our preliminary experiments showed that this approach alone is insufficient. For example, transferring from a single source speaker or speakers with similar timbre offers little to no improvement. Therefore, we perform speaker normalization on features correlated with timbre, such as unnormalized f0, to ensure that the matching criteria are not influenced by timbre-related aspects. This approach ensures that the unit selection algorithm generates references by selecting source speakers randomly and without bias.

We will then compare the ability of our prosodic-based model and gt-based model to preserve the target speaker's timbre when inferred using shuffle references with different speakers from the target. This comparison will include objective metrics for speaker similarity and subjective evaluations from both my personal and formal listening tests.

Personal Listening Test. When comparing the gt-based and prosodic-based models, it becomes evident that the prosodic-based model more effectively preserves the target speaker's timbre. This is particularly noticeable when the source and target speakers have distinctly different timbres, such as a male bass and a female soprano. The GT-based model often synthesizes a gender-neutral voice with significant noise and poor quality, whereas our model successfully replicates the female voice with stable performance. Although some speaker leakage still occurs, we suspect this is due to the limited number of speakers in the ESD dataset (only 10 in total) used for transfer. However, the leakage is much less frequent and generally less pronounced than with the GT-based model, indicating our model's ability to transfer speaker-independent prosodically-related features.

Speaker Similarity. To evaluate the models' ability to preserve speaker characteristics, we use the *WavLM-Large* model fine-tuned for the speaker verification task [32]. This model extracts speaker embeddings from both the reference and the utterance, then computes cosine similarity to assess timbre similarity [29]. We compare the performance of gt-based and prosodic-based models when inferred using non-target references from speakers different from the target. Additionally, we compute the similarity between two different speech samples from the same speaker, labeled as "samespk," to establish the highest baseline. Conversely, the similarity between samples from different speakers, labeled as "diffspk," provides the lowest baseline. To ensure a fair comparison, the speech samples from target speakers across different settings are consistent, with each sample being randomly selected from the test set and differing in text from the synthesized output used for cosine similarity computation. The results in Table 6.4 indicate that both model versions exhibit some leakage of speaker identity.

However, the prosodic-based model shows a clear improvement in preserving speaker identity compared to the gt-based model, demonstrating the effectiveness of our model in maintaining speaker characteristics. We also anticipate that by incorporating a wider variety of speakers for prosody transfer, the speaker leakage problem can be further mitigated, bringing performance closer to the highest baseline.

Configuration	**Speaker Similarity**
gtPt-shuffleRef	0.331 ± 0.017
prosodicPt-shuffleRef	0.389 ± 0.017
samespk	0.448 ± 0.019
diffspk	0.123 ± 0.027

Table 6.4: Speaker Similarity Scores for Different Model Configurations and Baselines.

Listening Test. To evaluate speaker identity preservation, participants were asked to compare pairs of samples generated by the gt-based and prosodic-based models using non-target references. They rated each sample on a 100-point scale, focusing on how closely the generated voices matched the original speaker's timbre. Also, to maintain focus on speaker identity preservation without the influence of speech quality, we excluded any samples with very poor quality. The results in Table 6.5 show that the prosodic-based model achieves a significantly higher speaker similarity score of 56.68 compared to the gt-based model's score of 43.56. This improvement validates our approach, demonstrating that incorporating speaker-independent prosodic features during training enhances the model's ability to disentangle timbre and effectively mitigate speaker leakage problem.

Configuration	**Speaker Similarity**
gtPt-shuffleRef	43.56 ± 7.13
prosodicPt-shuffleRef	56.68 ± 6.70

Table 6.5: Speaker Similarity Scores for Different Configurations

In summary, our evaluation shows that by incorporating speaker normalization and using prosodically-aligned references free of speaker identity cues, our approach effectively disentangles prosody from speaker characteristics, thereby mitigating the

speaker leakage problem. Additionally, we anticipate that introducing a more diverse range of speakers to transfer from during training will further enhance the model's ability to maintain consistent speaker identities.

Chapter 7

Conclusions

In this project, we addressed key challenges in traditional prosody transfer tasks, particularly the issues associated with the teacher-forcing training strategy that relied on ground truth references. This conventional approach often resulted in the unintended transfer of entangled features like speaker identity and content information, leading to significant performance degradation and speaker leakage during inference when non-target references were used.

To address these issues, we proposed using non-target, prosodically-aligned speech as references during training. This approach minimized the mismatch between training and inference, ensuring consistent performance even when the reference differed from the target. By generating speech through unit selection that minimized overall prosodic feature distances to the target, we created references that more accurately captured the target prosody, achieving transfer performance comparable to the gt-based method. Additionally, our method improved speaker identity preservation over traditional prosody transfer models by using references that were exclusively prosodically-informative, avoiding same-speaker references, and normalizing speaker-related features like f0 during unit distance computation. Furthermore, we explored modeling prosody at different granularities and found that complementing GST with LST led to more robust performance and nuanced prosody transfer.

7.1 Discussion and Future Work

Our project utilizes the FastSpeech2 model and the ESD dataset to investigate prosody transfer in speech synthesis. Although the quality of the generated speech may not reach the high fidelity achieved by recent large-scale models trained on extensive datasets,

our approach provides a robust platform for testing hypotheses related to prosody transfer. The framework we've developed is both flexible and modular, allowing for the generation of non-existent references that can be adapted to a wide range of reference-based transfer tasks, including style transfer, emotion embedding, and speaker adaptation. Through the careful design of unit selection criteria that target specific features like emotion, accent, or prosodic variations, we are able to create customized references that are closely aligned with the desired characteristics of the target speech. This adaptability ensures that our approach meets the unique demands of various tasks. Furthermore, the modular nature of our solution facilitates its integration into existing TTS pipelines beyond FastSpeech2, offering a valuable resource for researchers tackling similar challenges in the field of speech synthesis.

Future research could focus on refining the unit selection process and expanding the range of speech characteristics considered, to enhance the method's applicability. Additionally, applying this approach to more complex domains, like cross-lingual or multilingual TTS, or incorporating background audio, could provide new insights and extend its use beyond traditional speech synthesis. Another promising direction is integrating our reference generation framework with large-scale pre-trained models, which could improve controllability and expressiveness. Ultimately, by advancing prosody control in TTS, this research aims to develop more expressive and adaptable speech synthesis systems with significant implications for applications such as personalized speech synthesis and virtual assistants.

Bibliography

[1] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1448–1460, 2022.

[2] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pages 373–376. IEEE, 1996.

[3] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.

[4] Nathalie J Veenendaal, Margriet A Groen, and Ludo Verhoeven. The role of speech prosody and text reading prosody in children's reading comprehension. *British Journal of Educational Psychology*, 84(4):521–536, 2014.

[5] Delphine Dahan. Prosody and language comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(5):441–452, 2015.

[6] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[7] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[8] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[9] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman. Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech. *arXiv preprint arXiv:2004.14617*, 2020.

[10] Atli Thor Sigurgeirsson and Simon King. Do prosody transfer models transfer prosody$f$. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[11] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.

[12] Keith Ito and Linda Johnson. The lj speech dataset. 2017.

[13] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

[14] Steven R Livingstone and Frank A Russo. 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(5):e0196391, 2018.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[16] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2021.

[17] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.

[19] Douglas O'Shaughnessy. Design of a real-time french text-to-speech system. *Speech Communication*, 3(3):233–243, 1984.

[20] Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological bulletin*, 97(3):412, 1985.

[21] Carlos Gussenhoven. The phonology of tone and intonation. 2004.

[22] Francine R Chen and Margaret Withgott. The use of emphasis to automatically summarize a spoken discourse. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 229–232. IEEE, 1992.

[23] George D. Allen. Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3(2):75–86, 1975.

[24] Deborah A Hall and Christopher J Plack. Pitch processing sites in the human auditory brain. *Cerebral cortex*, 19(3):576–585, 2009.

[25] Yi Ren, Ming Lei, Zhiying Huang, Shiliang Zhang, Qian Chen, Zhijie Yan, and Zhou Zhao. Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7577–7581. IEEE, 2022.

[26] Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*, 2023.

[27] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[28] Martin Lenglet, Olivier Perrotin, and Gérard Bailly. Local style tokens: Fine-grained prosodic representations for tts expressive control. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 120–126. ISCA, 2023.

[29] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[30] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[32] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE, 2022.

Appendix A

Combined Participant Information Sheet and Consent Form

Our listening test involves human participants, and therefore requires their informed consent. The combined Participant Information Sheet and Consent Form are provided below.

Participant Information Sheet

Project title:	Advancing Prosody Transfer in Text-to-Speech with Pseudo Prosodic-similar Reference Audios
Principal investigator:	Simon King
Researcher collecting data:	Lin Liu
Funder (if applicable):	

Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?

The research team consists of Lin Liu, an MSc student at the University of Edinburgh, and Simon King, who is her supervisor and a Professor at the University of Edinburgh. Both of them will have access to the data to conduct and supervise the study.

What is the purpose of the study?

The purpose of this study is to compare the performance of prosody transfer between models trained with ground truth audio and those trained with prosodically similar but different audio. By analyzing participants' preferences and ratings of speech samples, we aim to evaluate and improve the effectiveness of these training methods

Why have I been asked to take part?

You have been asked to take part because you are a native English speaker with normal hearing ability, and are interested in listening tests. Your input will help us understand how different audiences perceive the prosody transfer performance of our models.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or

presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

What will happen if I decide to take part?

If you decide to take part, you will be asked to participate in a listening test. Here are the details:

- We will collect your ratings and preferences for various speech samples. This includes your opinions on the naturalness, clarity, timbre and prosody of the audio samples.
- Data will be collected through an online questionnaire where you will listen to speech samples and provide your feedback.
- Each listening test will last approximately 30 minutes
- You will complete one online session at a time and place convenient for you, using your computer or mobile device. The test is available at your convenience.

You will be paid £6 for your participation in this study.

Are there any risks associated with taking part?

There are no significant risks associated with participation.

Are there any benefits associated with taking part?

There are no direct benefits to you for taking part in this study. However, you will receive a small monetary reward as a token of appreciation for your participation.

What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of two years.

Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be

referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher/research team, including Lin Liu and Simon King.

All electronic data will be stored on a password-protected encrypted computer, or on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint). Your consent information will be kept separately from your responses in order to minimise risk.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. Please note that accessing your data may affect the outcome of the study. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

For general information about how we use your data, go to: edin.ac/privacy-research

Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Lin Liu, at s2491723@ed.ac.uk or her supervisor, Prof Simon King, at Simon.King@ed.ac.uk

This project has been approved by PPLS Ethics committee. If you have questions or comments regarding your rights as a participant, they can be contacted at 0131 650 4020 or ppls.ethics@ed.ac.uk.

Consent

By proceeding with the study, I agree to all of the following statements:

- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations.
- I allow my data to be used in future ethically approved research.