university of groningen

campus fryslân

Cross-lingual Voice Conversion and Its Prosodic Impact on Perceived Naturalness

Hao-Wei Liang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Cross-lingual Voice Conversion and Its Prosodic Impact on Perceived Naturalness

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)

Hao-Wei Liang (S-5962080)

June 11, 2025

Acknowledgements

First, I am deeply grateful for the guidance of my supervisor Dr. V. (Vass) Verkhodanova. She alerted me to the difficulties and limited viability of my initial research topic. After I changed to a new topic, her consistent support proved indispensable, she truly gave me the greatest freedom to explore my desired directions, directing the overall structure of this thesis and ultimately facilitating its completion.

I would also like to thank the faculty and staff of the Voice Technology program at the University of Groningen, whose teaching and resources laid the foundation for this work. Special thanks are due to the Center for Information Technology at the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster, which allowed me to train and experiment with the model.

I am also deeply grateful to my classmates and friends in the voice technology program, especially those who offered helpful discussions and emotional support. Even taking me out to go for a walk or giving me chocolate to get refreshed ideas and a sugar high means a lot to me.

Finally, I would like to thank my friends and family for their support and encouragement from far away in Taiwan. The survey would not have been possible without their patience, and the research would not have been completed.

Abstract

This thesis presents an exploratory investigation into the role of prosodic control in cross-lingual voice conversion between Taiwanese Mandarin and American English. As multilingual communication becomes more common in speech interfaces, language learning, and accessibility technologies, producing speech that sounds natural across language boundaries is a growing area of interest. However, the influence of prosodic features, particularly pitch and energy, on perceived naturalness in cross-lingual synthesis remains relatively underexplored, especially between typologically distinct languages such as tonal and non-tonal systems.

To explore this question, a FastSpeech2-based voice conversion model was trained using two open-source corpora: a subset of the Common Voice corpus containing Taiwanese Mandarin and the subset of the LJSpeech corpus containing American English. The two datasets were combined and used to train a single multilingual model. During inference, prosodic features were controlled under four conditions: baseline (no adjustment), pitch-only, energy-only, and combined pitch and energy control. The goal was to assess how these adjustments affect the perceived naturalness of synthesized speech.

A subjective listening test with 50 participants was conducted, in which each version was rated using a 5-point Likert scale. The results showed that the baseline condition consistently received the highest naturalness scores, while prosody, controlled versions, particularly the combined condition, were rated quite lower. This suggests that naive prosodic manipulation, without linguistic adaptation, may negatively affect the fluency and perceived coherence of synthesized cross-lingual speech.

To confirm that prosodic changes were successfully implemented, average pitch (F0) and RMS energy were extracted and compared across versions. Additionally, automatic speech recognition (ASR) metrics such as character error rate (CER) and word error rate (WER) were calculated as supplementary indicators of acoustic robustness. These scores are not intended to reflect human intelligibility, but rather to observe how prosody scaling affects system-level recognition.

This study offers initial insights into the limitations of uniform prosody control in cross-lingual voice conversion. The findings suggest that context-aware, linguistically informed prosody strategies may be needed to improve naturalness when converting between typologically diverse languages.
Keywords: Voice Conversion, Cross-lingual, FastSpeech2, Prosody Control, Perceived Naturalness, Pitch and Energy

Contents

1 Introduction

Voice conversion is a technique that adapts the perceived identity of a speaker in an utterance while preserving the linguistic content. It has attracted greater interest in recent decades because of its wide applicability in personalized speech synthesis, entertainment (e.g., voice dubbing), speaker anonymization, and assistive technologies for people with speech impairments (Sisman, Yamagishi, King, & Li, 2020). The purpose of voice conversion is not to change what is being said but how it is said, especially to change speakers' timbre or characteristics so that the speech sounds like it is spoken by someone else.

The development of voice conversion has evolved significantly over the past two decades. Early approaches were dominated by statistical methods such as Gaussian Mixture Models (GMMs) and frequency warping techniques, which relied heavily on parallel corpora and often produced over-smoothed and unnatural results (Stylianou, Cappé, & Moulines, 1998; Toda, Black, & Tokuda, 2007). With the growth of deep learning, it has led the voice conversion field to a new page, neural architectures like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and sequence-to-sequence (seq2seq) models that allowed non-parallel training and more flexible voice mappings (Hsu, Zhang, & Glass, 2017; Kaneko, Takaki, Kameoka, & Yamagishi, 2017). These models have outstandingly improved the naturalness and robustness of voice conversion systems through various applications.

Recent research has increasingly emphasized the goals of higher expressiveness, greater controllability, and broader language generalization in voice conversion systems. Several studies have supported this trend by incorporating prosodic features, particularly pitch and energy, into both training and inference processes to improve the realism of synthesized speech. For instance, Byun, Moon, and Visser (2023) demonstrated that frame-level conditioning on pitch and energy in a diffusion-based voice conversion model significantly enhanced intelligibility and voice quality. Similarly, Chen and Duan (2022) proposed ControlVC, a zero-shot voice conversion framework that introduced time-varying controls on pitch and speed, resulting in more dynamic and natural-sounding voice conversion outputs. These findings suggest that modeling prosody at a granular level contributes substantially to the naturalness and emotional expressiveness of converted speech.

Simultaneously, there is mounting interest in multilingual and cross-lingual voice conversion due to its potential real-world applications and technological advantages. Rather than limiting voice conversion systems to a single language, training on multilingual datasets enables models to generalize across a wider range of phonetic and prosodic patterns. This has been shown to be particularly beneficial for applications involving low-resource languages, where synthetic speech can serve as an effective tool for data augmentation and downstream model training(Baas & Kamper, 2021).

Such systems are also increasingly relevant in global communication contexts. In multilingual meetings or international conferences, generating synthesized speech with language-specific or familiar prosodic patterns may improve listener comprehension and engagement. For instance, Georgiou (2024); Smith, Holmes-Elliott, Pettinato, and Knight (2014) demonstrated that accent familiarity plays a significant role in speech perception, suggesting that accent-aware voice conversion could reduce cognitive load and increase accessibility in cross-linguistic settings.

In addition, multilingual voice conversion systems offer potential in language education, voice-assisted translation, and cross-cultural media production, where speech needs to sound natural and intelligible in multiple languages. The ability to generate speech that reflects the phonetic traits of multiple languages, while simultaneously maintaining prosodic naturalness, represents an important

step toward more inclusive and effective speech technologies.

1.1 Scientific Motivation

Recent developments in voice conversion have shown remarkable capabilities in transforming speech across speakers under monolingual conditions. Nonetheless, the exploration of prosody in cross-lingual voice conversion, especially its influence on perceived naturalness, remains limited. Prosody, including pitch, intensity, and duration, is essential to conveying natural-sounding speech and directly shapes how listeners perceive converted voices across different languages.

Du, Zhou, Sisman, and Li (2020) introduced a CycleGAN-based cross-lingual voice conversion framework that modeled pitch trajectories using continuous wavelet transform. Their work demonstrated improved subjective evaluations when prosody was explicitly modeled, compared to spectrum-only systems. They explicitly described their method as the first study addressing prosody in cross-lingual voice conversion. Zhao, Wang, Nguyen, and Ma (2021) also addressed prosodic mismatches by applying log-scale pitch normalization within a neural voice conversion pipeline. They reported that this prosodic control significantly improved the perceived naturalness of converted speech.

This study focuses on voice conversion between Taiwanese Mandarin and American English. These two languages differ in pitch usage, rhythmic structure, and stress assignment. Taiwanese Mandarin uses pitch to distinguish lexical tones, while English encodes meaning through intonation and stress placement. These differences can introduce challenges in prosodic transfer. Converted speech may appear unnatural or disfluent even if phonetic accuracy is achieved.

In practical applications such as multilingual voice assistants, language learning tools, or cross-lingual accessibility systems, the perceived naturalness of speech affects listener engagement and comprehension. Therefore, improving naturalness through prosodic control is both technologically and socially important.

To address these challenges, this study applies a FastSpeech2-based voice conversion system with pitch and energy control features during inference. Previous studies have shown that prosodic conditioning enhances expressiveness and subjective speech quality in voice conversion systems (Byun et al., 2023; Chen & Duan, 2022; Wang, Han, Lv, Zhou, & Chu, 2025). Building on this foundation, the current work aims to investigate how prosodic conditioning influences the perceived naturalness of cross-lingual speech, especially when converting between typologically different languages.

1.2 Social Motivation

Multilingualism is the norm rather than the exception. According to Di Pisa, Pereira Soares, and Rothman (2021), "over half of the world's population is at least bilingual, if not multilingual." Countries such as India and Singapore are well-known bilingual societies, where individuals commonly use different languages across education, work, and social domains. In such environments, the effectiveness of spoken communication depends not only on linguistic accuracy but also on how familiar the speech sounds to the listener.

In multilingual and multicultural communication, listeners tend to comprehend and relate more easily to speech when it aligns with familiar accentual or prosodic patterns. Georgiou (2024) found that listeners were more successful at understanding speech when it was delivered in an accent they were familiar with. Similarly, Smith et al. (2014) showed that long term and short term exposure

to an accent improves comprehension, even under acoustically challenging conditions such as background noise. These findings underscore the importance of accent familiarity in supporting speech intelligibility and reducing processing effort.

Voice conversion systems that generate speech with accentual patterns familiar to the listener, such as American accented Mandarin or Taiwanese accented English, have the potential to enhance communication across language boundaries. By aligning prosody with listener expectations, such systems can reduce cognitive load and increase speech clarity. This benefit is especially important in contexts such as cross cultural education, multilingual user interfaces, and accessibility technologies that serve linguistically diverse users.

The social motivation for this study is therefore closely connected to its scientific goal. Just as prosody influences perceived naturalness in cross lingual synthesis, it also determines how socially and cognitively accessible speech is to the listener. The ability to generate prosodically appropriate and familiar sounding speech can bridge gaps in cross linguistic interaction and improve real world usability of voice technologies.

1.3 Research Questions and Hypothesis

This study investigates how prosodic conditioning during inference affects the perceived naturalness of cross-lingual voice conversion between Taiwanese Mandarin and American English. The focus is placed on the use of pitch and energy modifications in a FastSpeech2-based voice conversion system.

> **How does adjusting pitch and energy during inference affect the perceived naturalness of cross-lingual speech produced by a FastSpeech2-based voice conversion model?**

This primary question leads to two subquestions:

1. How does modifying pitch at inference time influence listener ratings of naturalness in converted speech?

2. Do the perceptual effects of pitch and energy adjustment differ depending on the direction of conversion (Mandarin to English vs. English to Mandarin)?

Based on the research question, the hypothesis of the study is as below:
It is hypothesized that adjusting pitch and energy values during the inference stage of a FastSpeech2-based voice conversion system can improve the perceived naturalness of cross-lingual speech(Ren et al., 2020). This expectation is based on the linguistic role of prosodic features such as pitch and energy, which contribute to the rhythmic and intonational structure of speech. In tonal languages like Mandarin, pitch conveys lexical information that distinguishes words in the lexicon (Xu, 2005), while in English it serves to mark stress and intonation. By modifying these features to better align with the prosodic norms of the target language, the converted speech may sound more natural to listeners.

In this study, pitch and energy are not included as input features during training. Instead, fixed or manually adjusted values are applied to the variance adaptor module at inference time to manipulate prosodic characteristics. The study evaluates how these inference-time adjustments influence subjective ratings of naturalness in synthesized speech.

Now that the motivation for this research has been presented, the structure of this thesis is as follows: The structure of this thesis is as follows. Chapter 1 introduces the scientific and social motivations behind the study and outlines the research questions and Hypothesis. Chapter 2 provides a literature review covering five key areas: neural speech synthesis and FastSpeech 2, voice conversion with prosodic and cross-lingual focus, cross-lingual voice conversion with Mandarin-English focus, and prosodic features in voice conversion. Chapter 3 details the methodology, including dataset description, model architecture, alignment procedures, data preprocessing and prosodic feature extraction, cross-lingual training strategy and lexicon integration, participants and sampling, and ethical considerations. Chapter 4 presents the experimental setup, specifying the training process and evaluation methods. Chapter 5 presents the listner background and reports the results from both subjective and objective evaluations. Chapter 6 discusses the findings in relation to the research questions and hypothesis and also addresses limitations. Finally, Chapter 7 concludes the thesis and directs future work.

2 Literature Review

This chapter reviews existing literature relevant to neural speech synthesis, voice conversion, prosodic features, and cross-lingual voice conversion. The review is organized into the following sections: 2.1 Neural Speech Synthesis and FastSpeech2, 2.2 voice conversion with prosodic and cross-lingual focus, 2.3 cross-lingual voice conversion with Mandarin-English focus, and 2.4 prosodic features in voice conversion.

2.1 Neural Speech Synthesis and FastSpeech 2

Neural speech synthesis has undergone significant advancements over the past decade, transitioning from autoregressive architectures to more efficient non-autoregressive models. One of the earliest breakthroughs was WaveNet van den Oord et al. (2016), a deep autoregressive generative model capable of producing high-fidelity raw audio waveforms. Although WaveNet demonstrated remarkable improvements in naturalness over conventional parametric or concatenative systems, its sequential generation process limited its real-time applicability due to high computational costs.

To address the latency issues of autoregressive waveform generation, subsequent works explored text-to-spectrogram models such as Tacotron (Wang et al., 2017) and Tacotron 2 (Shen et al., 2018). These models introduced attention-based sequence-to-sequence architectures that could generate mel-spectrograms from text, which were then fed into vocoders like WaveNet. Tacotron-based models significantly improved the naturalness and intelligibility of synthesized speech by learning internal alignments between graphemes and acoustic features. However, they were still autoregressive and prone to alignment errors, irregular prosody, and mispronunciations, particularly when dealing with long or noisy inputs.

To overcome these limitations, non-autoregressive models began to emerge. FastSpeech Ren et al. (2019) introduced a fully parallel architecture that decoupled duration prediction from acoustic modeling, enabling faster and more stable synthesis. Instead of relying on attention mechanisms for alignment, FastSpeech utilized duration information extracted from an external aligner to model temporal structure explicitly. This allowed for a significant speed-up in inference and improved robustness compared to Tacotron.

FastSpeech 2 (Ren et al., 2020) improves upon its predecessor FastSpeech (Ren et al., 2019) by incorporating variance predictors for pitch, energy, and duration. This allows the model to explicitly learn and control prosodic features, leading to more expressive and natural-sounding speech. The inclusion of pitch and energy significantly enhances both the objective and subjective quality of synthesized speech. Furthermore, FastSpeech 2's parallel architecture ensures efficient training and inference, making it suitable for large-scale or multilingual applications

FastSpeech 2' s flexibility has led to widespread adoption in both monolingual and cross-lingual TTS research. Its architecture supports conditioning on speaker embeddings, language IDs, and phoneme-level inputs, which makes it adaptable to tasks involving speaker adaptation, code-switching, and voice conversion. Recent studies have also employed FastSpeech 2 as a backbone for prosody-sensitive synthesis and cross-lingual experiments, leveraging its controllable structure to explore accent transfer and pitch manipulation across languages.

The development from autoregressive models like Tacotron to non-autoregressive models like FastSpeech 2 represents a major shift in the field of neural speech synthesis. FastSpeech 2 strikes a balance between speed, quality, and controllability, providing a robust foundation for the present

study, which investigates the role of pitch and energy in cross-lingual voice conversion between Taiwanese Mandarin and American English.

2.2 Voice Conversion with Prosodic and Cross-Lingual Focus

Voice conversion modifies acoustic properties like pitch, energy and duration, while maintaining linguistic content. Recent work has demonstrated that integrating prosodic features directly into models enhances naturalness and fluency of the output.

In low-resource scenarios, one-shot and few-shot voice conversion approaches become valuable. Xie, Yang, Lei, Xie, and Su (2022) proposed a self-supervised one-shot voice conversion framework using variational autoencoders. Their system includes a speaker-related pitch encoder that directly models pitch contours and intensity from a single utterance, enabling rapid adaptation with minimal data while preserving prosodic information.

Prosodic mismatches across languages further complicate cross-lingual voice conversion. Converting between closely related stress-timed, non-tonal languages like English and German is relatively straightforward. However, converting from English to tonal languages like Mandarin introduces greater complexity, because lexical pitch contours in Mandarin significantly alter word meaning. Liu, Wen, Lu, and Chen (2020) identified in bilingual low-resource TTS that tone errors and prosodic mismatches often occur when training with English-dominant data. Their approach to tone preservation and data augmentation achieved improved intelligibility and accent metrics.

Some methods address this by explicitly modeling tone. Zhao et al. (2021) proposed FastSpeech-VC, which adds normalized log-f0 to compensate for prosodic differences between English and Mandarin. Their cross-lingual conversion achieved naturalness comparable to a professional TTS baseline on English–Mandarin tasks.

Recent work in voice conversion also explores pitch conditioning and reference-based models. Zuo et al. (2025) introduced PFlow-VC, which discretizes pitch tokens and uses pitch-conditioned flow matching, leading to smoother pitch trajectories and improved prosody in cross-domain tasks. Zhang et al. (2024) proposed RefXVC, that enhances performance by comprehensively leveraging reference speech information. This includes extracting fine-grained speaker embeddings to capture timbre changes , and using a pronunciation matching network to relate pronunciation with timbre across languages. Additionally, RefXVC integrates a multi-reference encoding technique to enrich content information and capture the full range of a speaker's voice. To ensure natural prosody in the converted speech, the system also introduces normalized pitch.

These methods illustrate the importance of explicitly modeling prosody in cross-lingual voice conversion. They show that when converting to tonal languages, systems must include mechanisms to reconstruct pitch contours, such as pitch encoders, discrete tokens, tone supervision, and reference-based embeddings, to preserve intelligibility and natural flow in the absence of tonal cues in the source speech.

2.3 Cross-lingual Voice Conversion with Mandarin-English Focus

Cross-lingual voice conversion between typologically distinct languages, such as Mandarin Chinese and American English, presents unique challenges due to differences in prosodic systems, phonemic inventories, and linguistic structures. Mandarin is a tonal language in which pitch contours are integral to lexical meaning, whereas English uses intonation and stress primarily for pragmatic

or syntactic marking. These fundamental disparities require voice conversion systems to learn and transfer complex prosodic patterns across languages while preserving linguistic accuracy and output intelligibility.

Several studies have proposed methods to address the challenges of converting between Mandarin and English. Zhou, Tian, Yılmaz, Das, and Li (2019) introduced a modularized neural network with language-specific output layers, allowing the model to separate language-dependent and language-agnostic components. This architecture improved generalization to both languages by isolating their respective linguistic properties. Zhao et al. (2021) proposed a controllable voice conversion framework that uses phonetic posteriorgrams (PPGs) as linguistic representations. Their system, built on a TTS decoder, enabled prosody control and achieved improved naturalness in cross-lingual synthesis by decoupling phonetic content from acoustic realization.

Cycle-consistent adversarial learning has also shown promise in this area. Du et al. (2020) employed a CycleGAN-based approach to model spectral and prosodic features between Chinese and English without requiring parallel data. This method allowed for prosodic transformations such as pitch and duration shifts while maintaining the content structure, even under non-aligned training conditions.

Another notable development involves the use of multilingual or shared phoneme representations. These strategies enable cross-lingual systems to generalize across diverse phonological structures, especially when used in conjunction with prosody-aware architectures. In the FastSpeech 2 framework, prosodic features such as pitch and energy can be explicitly modeled through dedicated variance predictors. Ren et al. (2020) showed that including these features leads to significant improvements in speech naturalness and clarity, particularly in languages with tonal distinctions like Mandarin.

2.4 Prosodic Features in Voice Conversion

Prosodic features, such as pitch (F0), energy, and duration, are essential elements in achieving natural and intelligible synthesized speech. In voice conversion systems, accurate modeling of these features helps convey rhythm, emphasis, and emotional nuances, which contribute directly to the perceived quality and fluency of the converted output.

Ren et al. (2020) proposed the FastSpeech 2 architecture, which includes variance predictors for pitch, energy, and duration. This model allows the synthesis system to explicitly represent prosodic variations, leading to more natural-sounding output across a variety of languages. Their experiments demonstrated that the inclusion of prosodic predictors significantly improves the perceptual quality of synthesized speech, especially when compared to systems that rely solely on phoneme and duration information.

In cross-lingual voice conversion tasks, prosodic mismatches between the source and target languages present particular difficulties. Mandarin Chinese, for example, encodes lexical meaning through tonal contours, whereas English uses intonation and stress without lexical tone. Du et al. (2020) addressed this challenge by designing a system that separates spectral and prosodic modeling paths. Their use of continuous wavelet transform to model F0 enabled more precise reconstruction of tonal patterns, which contributed to better perceived naturalness in Mandarin speech generated from English input.

Prosody modeling has also been explored through representation learning and data augmentation. Sigurgeirsson and King (2023) examined the limitations of reference-based prosody transfer

models, concluding that many systems fail to properly disentangle prosodic information from linguistic content. This observation highlights the need for more controlled and interpretable prosody representations. In response to this limitation, Deng et al. (2023) proposed the PMVC framework, which augments training data and masks input segments to encourage the model to learn distinct prosodic features. Their evaluation showed that this strategy produces expressive speech with improved prosodic variation, even in the absence of text supervision.

Across both monolingual and cross-lingual tasks, these studies underscore the importance of treating prosodic features as core components of the speech signal, rather than as optional enhancements. Accurate modeling of pitch, energy, and duration enables systems to produce speech that not only sounds more natural but also more closely matches the prosodic expectations of different linguistic audiences.

3 Methodology

The methodology includes seven key components. First, Section 3.1 outlines the selection and preprocessing of two open-source speech corpora: Common Voice Taiwanese Mandarin and LJSpeech English, carefully matched for duration and demographic consistency. Section 3.2 presents the architecture and implementation of FastSpeech 2, highlighting its controllable prosody mechanisms. Section 3.3 details the use of Montreal Forced Aligner (MFA) to obtain phoneme-level alignments essential for training. Section 3.4 explains how pitch and energy features were extracted and modified for experimental synthesis. Section 3.5 introduces a cross-lingual training setup, including bilingual phoneme lexicon integration and speaker conditioning. Section 3.6 describes participants and evaluation metrics. Finally, Section 3.7 addresses ethical considerations related to data usage and participant privacy, and Section 3.8 provides the repository and execution.

3.1 Dataset Description

This study uses two open-source speech corpora: the Taiwanese Mandarin subset of Mozilla Common Voice and LJSpeech.

The Taiwanese Mandarin corpus is a multi-speaker dataset sourced from Mozilla Common Voice (Ardila et al., 2019). It contains speech samples from male and female speakers aged 20 to 60 years. For this study, recordings from various female speakers in the age thirties were selected to ensure a certain amount of demographic consistency. This subdataset consists of approximately 3,000 utterances, and the total length of audio is around 185 minutes. All audio files were originally sampled at 48 kHz and were afterward downsampled to 22.05 kHz for compatibility with the TTS model.

The English dataset corpus is LJSpeech (Ito & Johnson, 2017), a widely adopted single-speaker American English corpus spoken by a female speaker. While the full corpus contains over 13,000 utterances, only a subset of around 1,600 utterances (approximately 185 minutes) was randomly selected using Python code. This sub-selection ensures that the English corpus duration closely matches the Taiwanese Mandarin subset to assist the progress and balanced training. All selected samples were also resampled to 22.05 kHz for compatibility with the TTS model, which aligned with the Mandarin data.

This balanced data setup was designed to support a cross-lingual voice conversion system based on FastSpeech 2, enabling relevant comparisons between Taiwanese Mandarin to American English and American English to Taiwanese Mandarin conversion tasks.

3.2 FastSpeech 2 Architecture and Implementation

In this study, I employed FastSpeech 2 as the skeleton model for voice conversion because of its efficient, high-quality end-to-end and non-autoregressive architecture, which enables parallelized speech synthesis with controllable prosodic features. Originally proposed by Ren et al. (2020), FastSpeech 2 improves upon its predecessor by incorporating additional acoustic features such as pitch, energy, and duration, allowing for more expressive and natural-sounding synthesized speech. The model involves a phoneme encoder, variance adaptor, and mel-spectrogram decoder, all of which are optimized through a multi-task loss that includes mean squared error for duration, pitch, energy, and mel-spectrogram reconstruction.

To implement this system, I adopted the open-source implementation of FastSpeech 2 provided by Chien, Lin, Huang, Hsu, and Lee (2021), which is widely used in academic and industrial research due to its modular structure and high compatibility with multilingual corpora. This implementation was selected because it includes well-integrated support for pitch and energy extraction, as well as phoneme-level duration alignment using preprocessed TextGrid files, which are important for my voice conversion setup. All training, validation, and inference procedures were based on this implementation, with customized modules to accommodate cross-lingual phoneme-based input.

While the base model architecture was preserved, specific adjustments were made to support multilingual phoneme tokenization. For example, inference scripts were extended to allow explicit control of pitch and energy values, enabling systematic manipulation of these prosodic features during synthesis for experimental analysis.

3.3 Forced Alignment using Montreal Forced Aligner

To achieve frame-level phoneme alignments required for duration modeling in FastSpeech 2, I used the Montreal Forced Aligner (MFA) toolkit (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), a widely used alignment system that contains pronunciation dictionaries and acoustic models for generating precise temporal boundaries between phonemes and speech signals.

For the Taiwanese Mandarin part of the corpus, I used the official pretrained acoustic model and pronunciation dictionary provided by MFA under the name ”Mandarin Taiwan mfa” (McAuliffe & Sonderegger, 2024). The dictionary represents phoneme sequences in International Phonetic Alphabet (IPA) format with tone marks, specifically adapted for Taiwanese Mandarin. For example, a sentence such as ” 他說他晚上會回家吃飯” (He said he would go home for dinner in the evening) would be turned into: "tʰ a˧ ʂ w o˧ tʰ a˧ w a˩˧ n ʂ a˧ ŋ xʷ ej˨ xʷ ejˊ tɕ a˧ ʈʂʰ ʐˌ˧ f a˨ n" These detailed tonal and phonemic information is necessary for accurate prosody modeling in tonal languages like Mandarin.

For the English data, LJSpeech, I used the CMU dict-based English pronunciation dictionary provided by MFA, referred to its name as ”English MFA” (Gorman, Howell, & Wagner, 2011). This dictionary uses the ARPAbet phoneme format and is compatible with American English corpora such as LibriSpeech and LJSpeech. The corresponding pretrained acoustic model ”English MFA” was used to align phoneme sequences to audio.

The result of TextGrid files from MFA was used to extract phoneme-level duration labels, which were directly fed into the FastSpeech 2 variance adaptor. These alignments are offered as an important supervisory signal for learning pitch, energy, and timing patterns across both Taiwanese Mandarin and American English domains.

3.4 Data Preprocessing and Prosodic Feature Extraction

To ensure the input consistency and quality required for cross-lingual voice conversion, a unified preprocessing pipeline was applied to both Taiwanese Mandarin and American English corpora. All raw waveforms were resampled to 22.05 kHz to match the FastSpeech 2 model configuration. Text transcriptions were converted into phoneme sequences using a custom lexicon (combined plus.dict), which merges tone-bearing IPA phonemes for Mandarin and ARPAbet phonemes for English. This lexicon provided a shared symbol space for bilingual modeling.

Phoneme durations were extracted through Montreal Forced Aligner (MFA) as detailed in Section 3.3. These durations aligned each phoneme to corresponding time spans in the waveform and served as input supervision for the variance adaptor in FastSpeech 2.

Fundamental frequency (F0) and root mean square (RMS) energy were extracted at the frame level without speaker-wise normalization or interpolation, preserving natural prosodic variation and allowing unvoiced regions to remain unaltered. These features were aligned with phoneme durations obtained from forced alignment and served as prosodic supervision during model training.

During inference, pitch and energy values were manipulated using fixed scalar multipliers, for example, doubling pitch or energy to simulate prosodic exaggeration. Four synthesis conditions were tested: baseline (1.0 pitch, 1.0 energy), pitch-scaled (2.0 pitch, 1.0 energy), energy-scaled (1.0 pitch, 2.0 energy), and both-scaled (2.0 pitch, 2.0 energy). This allowed for a systematic evaluation of how prosody affects perceived naturalness across languages.

Audio preprocessing was configured to match the model's architecture and vocoder requirements, including a sampling rate of 22.05 kHz, a short-time Fourier transform (STFT) frame size of 1024 with 256-hop spacing, and 80-dimensional mel-spectrograms spanning 0 to 8000 Hz. A subset of 512 utterances was reserved for validation throughout the training process.

3.5 Cross-lingual Training Strategy and Lexicon Integration

To facilitate cross-lingual voice conversion between Taiwanese Mandarin and American English, a multilingual FastSpeech 2 model was trained using a joint corpus. Unlike traditional monolingual synthesis systems, my approach unifies training data across languages and speakers within a single model. This strategy is essential for learning language-agnostic acoustic mappings and enables the model to generalize to mixed-lingual inputs.

Each utterance was processed into a phoneme sequence using the merged lexicon described in Section 3.4. For Mandarin data, phonemes were represented in tone-bearing IPA format using the Mandarin Taiwan MFA dictionary, while English phonemes followed the ARPAbet convention from the English MFA dictionary. To prevent symbol collision, each phoneme set was maintained under a disjoint namespace. For example, Mandarin symbols such as "t^h" and "$ʂ$" retained tonal markings, while English symbols like "T", "EH1", and "N" preserved their original ARPAbet encodings.

In addition to the phoneme sequence, each sample was tagged with both a speaker ID and an optional language indicator, enabling the model to implicitly learn language-specific prosodic patterns. During training, both phoneme and speaker/language embeddings were processed through a shared encoder-decoder structure. The variance adaptor handled duration, pitch, and energy, while mel-spectrograms served as the final output target. The system was trained using a multi-task loss function that jointly optimized pitch, energy, duration, and mel-spectrogram reconstruction.

This multilingual training framework enabled inference scenarios where Mandarin speech could be synthesized using an English speaker's voice characteristics, and vice versa. Furthermore, the inclusion of pitch and energy as controllable variables opened up a wide range of expressive possibilities, supporting the core experimental analysis in this study.

3.6 Participants and Sampling

In addition to training a FastSpeech 2-based voice conversion system, this study also incorporated a subjective evaluation component to assess the naturalness of the synthesized speech. A listening test

was designed in which participants compared utterances generated under different prosodic settings. Specifically, pitch (F0) and energy (RMS) features were manipulated during inference to observe how these prosodic variations influenced perceived naturalness across cross-lingual conversion conditions.

While a detailed explanation of the evaluation stimuli and test design is provided in Chapter 4, this section introduces the procedures surrounding participant recruitment and the ethical measures taken during the listening experiments.

Participants were recruited using convenience sampling, primarily through social and academic networks. A total of 50 valid responses were collected via an online questionnaire platform (Qualtrics XM). The survey did not collect personal identifiable information such as names, contact details, or IP addresses in the final dataset. Demographic data collection was limited to language background.

Specifically, participants were asked to indicate their native language and rate their self-perceived comprehension ability in both Mandarin Chinese and English using a 5-point Likert scale. This information was used to ensure that the listeners had sufficient understanding of the target languages to evaluate the synthesized utterances reliably.

The use of convenience sampling, while common in perceptual evaluation studies, limits the generalizability of the results. However, the collected linguistic background data allows for basic interpretive contextualization of the ratings, especially in cross-linguistic comparisons.

3.7 Ethical Considerations

This study adhered to the ethical guidelines of the University of Groningen and complied with the General Data Protection Regulation (GDPR). All participants were informed about the purpose, procedures, and voluntary nature of the study via an information and consent statement presented at the beginning of the online questionnaire. No personally identifiable information such as names or contact details was collected. Responses were anonymized, and data were stored securely in an institutional research drive. The questionnaire did not involve vulnerable populations and posed minimal risk. Participants were allowed to withdraw at any point without consequences. All ethical considerations were documented in the accompanying ethics application.

3.8 Repository and Execution

All code developed for this study is publicly available at the following GitHub repository: `https://github.com/LydonLiang/Thesis_VC` Detailed setup and usage instructions are provided in the README.md file. The dataset used in this study is not included in the repository and must be downloaded separately from the official Common Voice and LJSpeech sources.

4 Experimental Setup

This chapter outlines the experimental configuration used to implement and evaluate the proposed cross-lingual voice conversion system. Section 4.1 describes the architectural parameters, training procedure, and computing resources involved in training a FastSpeech 2 model capable of prosody-conditioned synthesis. Section 4.2 details the evaluation framework, combining subjective listening tests with objective acoustic and ASR-based analysis. The evaluation focuses on measuring how different prosodic manipulations, specifically pitch and energy scaling applied during inference, affect the perceived naturalness of synthesized speech across Mandarin and English utterances.

4.1 Training Setup

The training process for the voice conversion system was conducted using the FastSpeech 2 architecture based on the open-source implementation provided by Chien et al. (2021). This variant supports multi-speaker modeling and prosody control, making it well-suited for cross-lingual voice conversion tasks.

The model employs 4 encoder layers and 6 decoder layers, each with 2 attention heads and a hidden size of 256 units. The convolutional filter size in the transformer' s feed-forward blocks was set to 1024, with kernel sizes of [9, 1], and a dropout rate of 0.2 was applied in both encoder and decoder layers. The variance adaptor includes dedicated pitch and energy predictors, each using a convolutional filter size of 256 and a kernel size of 3. Linear quantization was applied to both pitch and energy with 256 bins, and all prosodic features were extracted and modeled at the phoneme level.

The system was trained using a batch size of 16 on 8 NVIDIA A100 GPUs provided by the Hábrók high-performance computing cluster. The optimizer was Adam with betas of (0.9, 0.98), an epsilon of 1e-9, and no weight decay. A learning rate schedule with 4000 warm-up steps and exponential decay was used, with annealing points set at 300k, 400k, and 500k steps. The training process was configured to run for a total of 50,000 steps, with validation and synthesis performed every 1,000 steps, and checkpoints saved every 10,000 steps. For waveform reconstruction, HiFi-GAN (universal version) was used as the neural vocoder. This vocoder was selected for its efficiency and strong perceptual quality, ensuring that the generated waveforms remained natural and expressive across both source and target languages.

The training included both in-domain and cross-lingual data: Taiwanese Mandarin from the Common Voice corpus and American English from the LJSpeech dataset. Each language was associated with a single representative speaker, and speaker embeddings were enabled to distinguish between them.

Training and validation logs indicate a stable convergence pattern. For instance, the total validation loss dropped from 3.25 at step 1,000 to around 2.65–2.70 by step 10,000 and remained within a comparable range through step 20,000. Mel-spectrogram reconstruction losses consistently decreased from approximately 0.78 to 0.60, showing improvement in spectral fidelity. Pitch and energy losses stabilized around 1.0 and 0.36, respectively, while duration loss remained under 0.09, demonstrating that the model effectively learned to reproduce timing and prosodic features without severe overfitting.

Notably, pitch loss remained relatively high throughout training, a known challenge in non-parallel voice conversion settings where accurate pitch modeling is more difficult due to speaker

mismatch. Nonetheless, the pitch loss trend was relatively flat after the initial drop, suggesting stable modeling rather than collapse.

4.2 Evaluation

To evaluate the effects of prosodic features on the naturalness of cross-lingual voice conversion, this study adopts a comprehensive evaluation strategy involving both subjective listening tests and objective acoustic and ASR-based analyses. The primary goal is to determine how inference-time pitch and energy manipulations influence the perceived naturalness of speech synthesized using a FastSpeech2-based voice conversion system between Taiwanese Mandarin and American English.

The selection of evaluation metrics was guided by the nature of the research questions. Since naturalness is inherently a perceptual quality, a subjective listening test using a 5-point Likert scale was employed to measure listener judgments. To complement this, an additional Likert-scale question was used to evaluate whether listeners could perceive differences in prosody or expressiveness across versions of the same utterance. These tests directly correspond to the study' s hypothesis regarding listener perception of pitch and energy variations.

For the subjective evaluation, participants were recruited via convenience sampling and asked to complete an online listening test. The test comprised eight sentences (four Mandarin and four English), each synthesized under four prosodic conditions:

group	pitch	energy	explain
A	1.0	1.0	baseline
B	2.0	1.0	increased pitch only
C	1.0	2.0	increased energy only
D	2.0	2.0	increased both

Table 1: Prosodic conditions

These numeric scaling values correspond to command-line inference-time parameters passed to the synthesis script. Internally, they operate by multiplying each value in the original pitch or energy vector by the specified scalar. For example, pitch=2.0 means that every frame-level F0 value extracted from the input utterance is doubled prior to being passed into the variance adaptor of the FastSpeech2 model. This does not involve changing the training process or learned parameters but alters the synthesized prosody by exaggerating or preserving the original contour.

For each sentence, participants rated the naturalness of each audio version on a 5-point scale. After listening to all four versions, they also rated how much prosodic or expressive difference they could perceive among them. The presentation order of the audio versions was randomized for each sentence to minimize order effects. A total of 50 valid responses were collected. Responses were normalized and reshaped into long-form format to enable group-level statistical analyses based on version, pitch level, energy level, and language.

On the objective side, mean F0 and RMS energy were selected to verify whether the intended prosodic manipulations were realized in the acoustic signal. These features are standard indicators of prosody in speech synthesis research and were extracted at the frame level and averaged across each utterance. While mean values do not capture temporal dynamics, they provide a concise summary

of pitch height and energy intensity per sentence, which supports comparison across versions and conditions. These summary statistics were selected to match the resolution of subjective judgments, which were collected per sentence.

In addition, a Whisper-based automatic speech recognition (ASR) system was used to transcribe the synthesized utterances. The resulting transcripts were compared with ground truth references to calculate character error rate (CER) and word error rate (WER). While these metrics do not directly measure intelligibility from a human perspective, they serve as indirect indicators of how prosodic variation might affect speech recognizability from a system's standpoint. As such, they provide complementary insights into how pitch and energy adjustments impact machine transcription robustness, which may loosely correlate with signal clarity but not with listener comprehension per sentence.

All subjective and objective measures were subsequently analyzed and compared. Correlation analysis was performed to determine whether acoustic measurements (F0, energy) and recognition scores (WER, CER) aligned with subjective naturalness ratings. In addition, results were segmented by language (Mandarin vs. English) to explore interaction effects between language type and prosodic manipulation. This multifaceted evaluation framework directly addresses the study's research questions by providing converging evidence on how pitch and energy manipulations affect both perception and signal-based characteristics in cross-lingual TTS systems.

5 Results

This chapter presents the results of both subjective and objective evaluations conducted to assess the impact of prosodic manipulation on cross-lingual voice conversion. Section 5.1 summarizes the linguistic background of participants who took part in the listening test. Section 5.2 reports the naturalness ratings across different prosodic conditions, supported by statistical analyses to examine perceptual differences. Section 5.3 visualizes these ratings using boxplots to illustrate distributional patterns. Finally, Section 5.4 provides objective metrics, including fundamental frequency (F0), RMS energy, and automatic speech recognition (ASR) error rates, to validate the consistency of prosodic modifications and their potential effects on signal quality.

5.1 Listener background Summary

A total of 50 valid responses were collected through an online questionnaire distributed via social media platforms targeting Taiwanese users. Due to ethical considerations, no vulnerable populations, such as children or elderly individuals were included, and no personally identifiable information was collected. Although the questionnaire did not include a specific age question, based on the typical demographic of the social channels used, the majority of participants were estimated to be between 25 and 50 years old.

All participants were native speakers of Mandarin Chinese, and the questionnaire included a self-assessment of language comprehension abilities in both Mandarin and English on a 5-point Likert scale. As expected, nearly all respondents reported high comprehension in Mandarin. A smaller portion indicated high comprehension in English, with only a few reaching the maximum score of 5. Familiarity with speech synthesis or speech technology was not assessed, and this remains a limitation in the current demographic profile.

5.2 Subjective Naturalness Rating Analysis

To assess how different prosodic conditions affected perceived naturalness, listener ratings for each of the eight sentences were analyzed individually. As each question utilized a 5-point Likert scale, the data was treated as ordinal, and thus medians were used instead of means to summarize central tendencies. The Friedman test, a non-parametric statistical test for related samples, was employed to determine whether the differences in ratings across the four conditions (A–D) were statistically significant.

As shown in Table 2, version A (baseline with default pitch and energy) consistently received the highest or near-highest median scores, while version D (increased pitch and energy) generally received the lowest. The Friedman test revealed significant differences in perceived naturalness across the four versions for all sentences ($p < .001$), except for Sentence 006, which showed a smaller yet still statistically significant difference ($p = .021$). These findings indicate that listeners were sensitive to prosodic modifications and generally preferred more moderate or unaltered prosodic contours.

Table 2 summarizes the median naturalness ratings per version and the corresponding Friedman test results for each sentence.

sentence id	A	B	C	D	Friedman χ^2	p-value
sent001	3.0	2.0	4.0	2.0	71.70	<.001
sent002	3.0	2.0	2.0	2.0	54.78	<.001
sent003	4.0	2.0	3.0	2.0	78.88	<.001
sent004	3.0	2.0	2.0	1.0	84.25	<.001
sent005	3.0	3.0	3.0	2.0	38.01	<.001
sent006	3.0	2.0	3.0	3.0	9.71	0.021
sent007	4.0	3.0	3.0	3.0	29.83	<.001
sent008	3.0	3.0	2.0	2.0	58.43	<.001

Table 2: Median naturalness ratings per version and the corresponding Friedman test results

5.3 Graphical Visualization

Detailed boxplots for each sentence are presented in Appendix B. These visualizations support the statistical findings and illustrate distributional differences across the four prosodic conditions.Boxplots showing naturalness ratings for all eight sentences under the four prosodic conditions (A–D). The boxplots visualize the distribution of listener ratings and highlight version-wise differences in perceived naturalness.

5.4 Oblective Evaluation Result

To supplement the subjective listening test, I conducted an objective evaluation focusing on two aspects: the realization of prosodic manipulations and the robustness of automatic transcription. Specifically, I analyzed each synthesized speech sample to verify whether the intended modifications in pitch and energy were correctly applied, and I assessed how these manipulations might affect automatic recognition performance.

The first analysis involved extracting the average F0 (fundamental frequency) and RMS energy from each utterance. These acoustic features were measured using frame-level analysis and averaged across the duration of each sentence. The results confirmed that the prosody control settings passed at inference time were reflected in the actual synthesized audio: samples with doubled pitch parameters exhibited substantially higher mean F0 values, and energy-scaled versions showed corresponding increases in RMS energy. This step was essential to ensure that the experimental manipulation of prosody was both effective and consistent across conditions.

In the second analysis, I employed the Whisper ASR system to transcribe each synthesized utterance. I then computed Word Error Rate (WER) and Character Error Rate (CER) by comparing the ASR outputs to the ground-truth text. These metrics are often used as proxies for acoustic clarity or system-level intelligibility in synthetic speech studies. The baseline condition (version A) consistently achieved lower error rates across both Mandarin and English samples, while the most exaggerated prosodic condition (version D) often produced higher WER and CER, particularly in Mandarin sentences. This suggests that excessive prosodic manipulation can negatively impact the reliability of automatic decoding.

However, it is important to emphasize that ASR metrics do not measure human perception. While they offer insight into how well a speech signal is preserved for machine recognition, they cannot substitute for human intelligibility judgments. Therefore, these findings should be interpreted as indicative of signal robustness rather than perceptual comprehensibility.

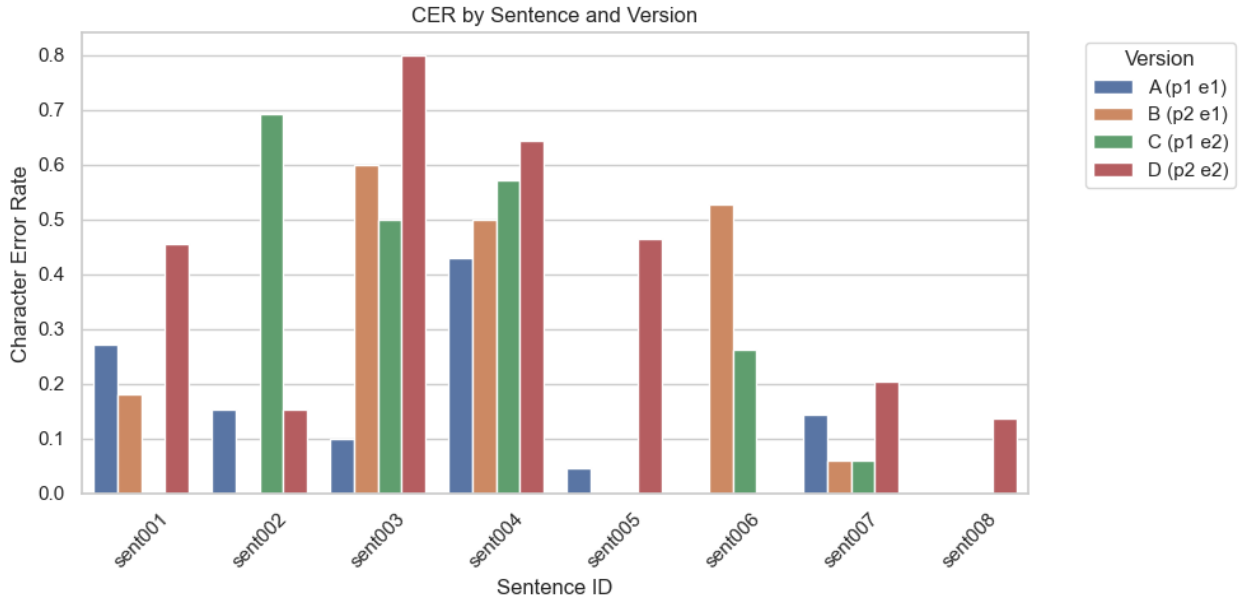


Figure 1: CER by sentence & version

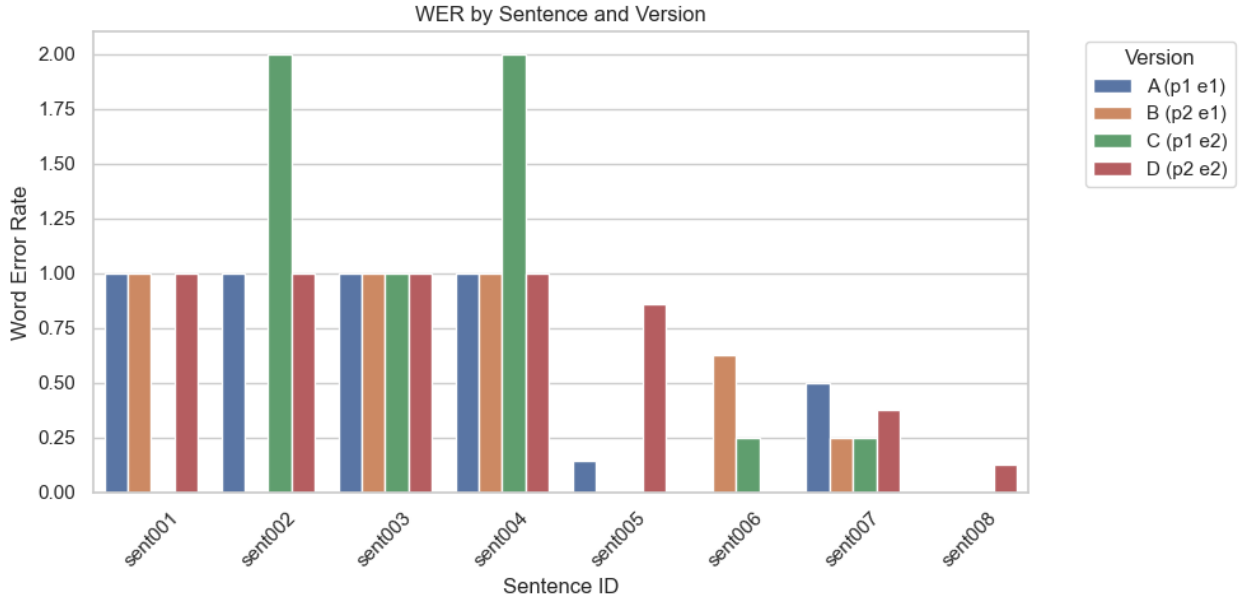


Figure 2: WER by sentence & version

6 Discussion

This chapter presents the results of the study in relation to the research questions and hypothesis outlined in Chapter 1. Section 6.1 addresses the main research question and its two sub-research questions by examining how pitch and energy adjustments affected listener judgments across different language directions. Section 6.2 evaluates whether the findings support the initial hypothesis that prosody manipulation would improve perceived naturalness in cross-lingual synthesis. Section 6.3 reflects on the methodological and modeling limitations encountered during the study and suggests directions for future research, particularly in terms of more context-aware and linguistically informed prosody control.

6.1 Answering RQ

This study aimed to examine how prosodic conditioning, specifically pitch and energy manipulation at inference time, affects the perceived naturalness of cross-lingual speech generated using a FastSpeech2-based voice conversion model. The findings from the subjective evaluation provide direct insight into the primary research question and its two subcomponents.

For the main research question, the results clearly indicate that adjusting pitch and energy during inference does have a measurable impact on listener-perceived naturalness. Across all eight sentences (four in Mandarin and four in English), the baseline version (condition A), which applied no prosodic scaling, consistently received the highest or near-highest naturalness ratings. In contrast, condition D, which involved simultaneously increasing both pitch and energy by a factor of two, resulted in significantly lower ratings. This suggests that while the FastSpeech2 architecture allows for prosody control, excessive modification of these parameters may degrade the naturalness of the output, likely due to unnatural rhythm, emphasis, or vocal effort artifacts introduced by the over-scaling.

Regarding sub-question 1, both pitch-only (condition B) and energy-only (condition C) manipulations showed intermediate effects, typically lowering perceived naturalness compared to the baseline but not to the extent seen in condition D. Statistical analysis (Friedman and Wilcoxon tests) confirmed that these differences were significant in most cases, reinforcing the sensitivity of listeners to even moderate prosodic variation in synthetic speech.

Sub-question 2 explored whether these effects differ depending on the conversion direction. The results suggest that conversions from Mandarin to English and from English to Mandarin exhibit asymmetric sensitivity to prosodic changes. In Mandarin sentences, which are tonal, exaggerated pitch or energy often led to steeper drops in naturalness, whereas English sentences showed slightly more tolerance. This highlights the linguistic constraints imposed by tonal systems and underscores the importance of language-aware prosody design in cross-lingual voice conversion.

In addition to subjective listener ratings, objective acoustic measurements were used to validate the implementation of prosodic manipulations. Frame-level analysis of F0 and energy confirmed that the intended adjustments were effectively realized in the synthesized output. Furthermore, automatic transcription results using Whisper ASR showed increased error rates, particularly in Mandarin sentences under the exaggerated prosody condition (D). While these ASR-based metrics do not directly reflect listener comprehension, they provide further evidence that extreme prosodic changes may introduce distortions or irregularities in the signal that affect decoding performance. Together, these subjective and objective findings suggest that prosody control must be applied with caution, especially in tonal languages.

6.2 Validation of the Hypothesis

This study hypothesized that modifying pitch and energy at inference time would enhance the perceived naturalness of cross-lingual speech produced by a FastSpeech2-based voice conversion system. The rationale was that prosodic features, particularly pitch and energy, are essential for conveying rhythm, emphasis, and intonation. Therefore, aligning these features more closely with the target language was expected to improve naturalness.

However, the results of the subjective evaluation did not support this hypothesis. The baseline version (A), which preserved the default prosody without any modifications, consistently received the highest naturalness scores across all evaluated sentences. In contrast, prosodically manipulated versions, particularly version D, which applied both pitch and energy scaling, were rated significantly lower. These differences were statistically confirmed and consistent across both Mandarin and English utterances. Contrary to the original hypothesis, exaggerated prosodic manipulation appeared to degrade rather than enhance perceived naturalness.

To further understand the causes of these results, a qualitative listening analysis was conducted by the researcher. This involved closely listening to the output samples across all prosodic conditions to detect audible irregularities that may not be captured in numeric scores. Several version D samples exhibited unnatural buzzing, excessive resonance, and occasional clipping at sentence endings. In some cases, syllables or final phonemes were dropped altogether. These artifacts were primarily observed in high-pitched voiced segments, suggesting a mismatch between the scaled prosody and the waveform generation module. This aligns with Skerry-Ryan et al. (2018)'s observation that their prosody representation encodes pitch in an absolute manner when transferring prosody from a reference signal, which can lead to the synthesized speech sounding unnaturally as if the target speaker is imitating a person with a substantially deeper or higher vocal range. Their findings suggest that speaker-dependent pitch content is transferred from the reference to the output , highlighting a potential entanglement of prosodic features with speaker identity that complicates harmonious synthesis across dissimilar speakers.

Furthermore, in tonal languages such as Mandarin, where pitch conveys lexical distinctions, inappropriately scaled pitch contours can interfere with both tone perception and sentence-level phrasing. Shen, Deutsch, and Le (2011) showed that elevating overall pitch height significantly reduced tone identification accuracy, especially for Tone 3, highlighting how excessive prosodic manipulation can introduce perceptual dissonance.

Objective ASR-based analysis further supported these observations: version D exhibited the highest average Character Error Rate (CER = 0.357) and Word Error Rate (WER = 0.670), compared to baseline and intermediate conditions. While CER/WER are not direct measures of human comprehension, they serve as meaningful proxies for signal clarity and decoding difficulty.

Together, these findings suggest that naïve prosody scaling, implemented without linguistic context or acoustic model adjustments may degrade both perceptual and acoustic quality in cross-lingual voice conversion. Future systems would benefit from context-aware prosody adaptation to preserve naturalness and signal stability.

6.3 Limitations

While this study provides valuable insights into the perceptual effects of prosody manipulation in cross-lingual voice conversion, several limitations must be acknowledged. First, the prosodic adjust-

ments explored in this work were unidirectional: both pitch and energy were only scaled upwards. Without testing downward modifications, it remains unclear whether reducing prosodic intensity might produce inverse effects or lead to similar degradation in naturalness.

Second, the use of a multi-speaker corpus for Taiwanese Mandarin (Common Voice) and a single-speaker corpus for English (LJSpeech) introduces variability in speaker representation. This asymmetry could have influenced listener perception and made it more difficult to isolate the effects of prosody alone. Additionally, the prosody control mechanism employed was based on global scaling across entire utterances. This uniform adjustment lacks contextual sensitivity and may have disrupted natural phrasing, particularly in languages like Mandarin where tonal contours interact with syntactic boundaries.

Lastly, the perceptual evaluation focused solely on naturalness ratings using a Likert scale. Other dimensions such as intelligibility, listener preference, or comprehension were not measured, limiting the overall interpretability of user experience. Future studies could address these issues through more fine-grained modeling and multi-dimensional evaluation approaches.

7 Conclusion

This thesis set out as an exploratory investigation into the role of prosody control in cross-lingual speech synthesis using a FastSpeech2-based voice conversion system. Specifically, it examined how inference-time manipulation of pitch and energy affects the perceived naturalness of synthesized utterances in both Mandarin and English. Although the original hypothesis predicted an improvement in naturalness through prosodic enhancement, the results consistently indicated the opposite: naturalness ratings were highest for baseline speech without any prosody scaling.

These findings suggest that naïve manipulation of prosodic parameters, without linguistic contextualization or phonological awareness, can disrupt the perceived coherence and fluidity of speech. Subjective ratings, auditory artifact inspection, and ASR-based error rates all pointed to a degradation of quality when pitch and energy were aggressively increased. The asymmetry between Mandarin and English responses further underscores the need to treat tonal and non-tonal languages differently in prosody design. Rather than generalizing prosody as a universal enhancement strategy, this study highlights the intricacies of cross-lingual rhythm, stress, and tone interactions.

As an exploratory study, this work did not aim to provide a fully optimized system, but rather to map out potential challenges and sensitivities that arise from prosody-driven modifications. The use of transparent, interpretable controls, though intentionally limited, helped isolate specific variables and provided a clear picture of their perceptual consequences.

7.1 Futurework

Building on these insights, future research could explore more context-sensitive forms of prosody manipulation. For example, prosody could be conditioned on syntactic boundaries, discourse cues, or semantic focus, rather than applying global scaling across entire utterances. This would allow synthesized speech to better reflect natural speech dynamics and avoid the artifacts associated with uniform scaling.

Additionally, data-driven approaches such as prosody prediction models trained on bilingual or code-switched corpora may help achieve more natural and adaptive prosodic contour shaping. The integration of language-specific phonological knowledge, such as tone sandhi patterns in Mandarin or stress-timed rhythm in English, could further improve the fluency and authenticity of generated speech across languages.

Beyond the modeling itself, future work could also enhance the evaluation framework. Expanding listener tests to include preference judgments, comprehension accuracy, or emotion appropriateness may provide a more holistic understanding of synthesis quality. As cross-lingual TTS systems continue to mature, bridging the gap between phonological theory and neural synthesis remains a promising and underexplored direction. This study contributes a first step by identifying perceptually sensitive variables and outlining concrete areas for refinement.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., … Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baas, M., & Kamper, H. (2021). Voice conversion can improve asr in very low-resource settings. *arXiv preprint arXiv:2111.02674*.

Byun, K., Moon, S., & Visser, E. (2023). Highly controllable diffusion-based any-to-any voice conversion model with frame-level prosody feature. *arXiv preprint arXiv:2309.03364*.

Chen, M., & Duan, Z. (2022). Controlvc: Zero-shot voice conversion with time-varying controls on pitch and speed. *arXiv preprint arXiv:2209.11866*.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., & Lee, H.-y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *Icassp 2021 - 2021 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 8588-8592). doi: 10.1109/ICASSP39728.2021.9413880

Deng, Y., Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023). Pmvc: Data augmentation-based prosody modeling for expressive voice conversion. In *Proceedings of the 31st acm international conference on multimedia* (pp. 184–192).

Di Pisa, G., Pereira Soares, S. M., & Rothman, J. (2021). Brain, mind and linguistic processing insights into the dynamic nature of bilingualism and its outcome effects. *Journal of Neurolinguistics*, *58*, 100965. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0911604420301251` doi: https://doi.org/10.1016/j.jneuroling.2020.100965

Du, Z., Zhou, K., Sisman, B., & Li, H. (2020). Spectrum and prosody conversion for cross-lingual voice conversion with cyclegan. In *2020 asia-pacific signal and information processing association annual summit and conference (apsipa asc)* (pp. 507–513).

Georgiou, G. P. (2024). Perception of familiar second language accents and the role of linguistic background. *Applied Sciences*, *14*(24), 11776.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Hsu, W.-N., Zhang, Y., & Glass, J. (2017). Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in neural information processing systems*, *30*.

Ito, K., & Johnson, L. (2017). The lj speech dataset.

Kaneko, T., Takaki, S., Kameoka, H., & Yamagishi, J. (2017). Generative adversarial network-based postfilter for stft spectrograms. In *Interspeech 2017* (pp. 3389–3393).

Liu, R., Wen, X., Lu, C., & Chen, X. (2020). Tone learning in low-resource bilingual tts. In *Interspeech* (pp. 2952–2956).

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (Vol. 2017, pp. 498–502).

McAuliffe, M., & Sonderegger, M. (2024, Feb). *Mandarin (taiwan) mfa dictionary v3.0.0* (Tech. Rep.). `https://mfa-models.readthedocs.io/pronunciationdictionary/Mandarin/Mandarin(Taiwan)MFAdictionaryv3_0_0.html`.

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. *arXiv preprint arXiv:2006.04558*.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, *32*.

Shen, J., Deutsch, D., & Le, J. (2011). The effect of overall pitch height on mandarin tone identifi-

cation. In *Proceedings of meetings on acoustics* (Vol. 14).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., … others (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4779–4783).

Sigurgeirsson, A. T., & King, S. (2023). Do prosody transfer models transfer prosody*f*. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 132–157.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., … Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning* (pp. 4693–4702).

Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R.-A. (2014). Cross-accent intelligibility of speech in noise: Long-term familiarity and short-term familiarization. *Quarterly Journal of Experimental Psychology*, *67*(3), 590–608.

Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, *6*(2), 131–142.

Toda, T., Black, A. W., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2222–2235.

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

Wang, Y., Han, X., Lv, S., Zhou, T., & Chu, Y. (2025). Mpfm-vc: A voice conversion algorithm based on multi-dimensional perception flow matching. *Applied Sciences*, *15*(10), 5503.

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

Xie, Q., Yang, S., Lei, Y., Xie, L., & Su, D. (2022). End-to-end voice conversion with information perturbation. In *2022 13th international symposium on chinese spoken language processing (iscslp)* (pp. 91–95).

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech communication*, *46*(3-4), 220–251.

Zhang, M., Zhou, Y., Ren, Y., Zhang, C., Yin, X., & Li, H. (2024). Refxvc: Cross-lingual voice conversion with enhanced reference leveraging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhao, S., Wang, H., Nguyen, T. H., & Ma, B. (2021). Towards natural and controllable cross-lingual voice conversion based on neural tts model and phonetic posteriorgram. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5969–5973).

Zhou, Y., Tian, X., Yılmaz, E., Das, R. K., & Li, H. (2019). A modularized neural network with language-specific output layers for cross-lingual voice conversion. In *2019 ieee automatic speech recognition and understanding workshop (asru)* (pp. 160–167).

Zuo, J., Ji, S., Fang, M., Jiang, Z., Cheng, X., Yang, Q., … others (2025). Enhancing expres-

sive voice conversion with discrete pitch-conditioned flow matching model. *arXiv preprint arXiv:2502.05471*.

Appendices

A Questionnaire Survey

The Role of Prosodic Features in Voice Conversion between Taiwanese Mandarin and American English
音韻特徵對臺灣中文與美式英語語音轉換自然度之影響研究

Dear participant,

Thank you for participating in this study. This research investigates the role of prosodic features, specifically pitch and energy, in a cross-lingual voice conversion text-to-speech (TTS) system between Taiwanese Mandarin and American English. The study aims to evaluate how different speech versions affect the perception of naturalness and prosodic variation.
You will listen to several sets of speech samples, each consisting of four versions. After each set, you will be asked to answer two short questions.
Estimated time to complete: 5–10 minutes No prior knowledge of linguistics is required All data will be recorded anonymously and used solely for academic research. No personally identifiable information will be collected.
Participation is entirely voluntary. You may withdraw at any time without any consequences. Your data will be processed in accordance with the General Data Protection Regulation (GDPR). You have the right to access, rectify, or erase your data at any point before anonymization.
If you have any questions, please contact the researcher:
Researcher: Hao-Wei Liang
MSc in Voice Technology, University of Groningen
Email: h.w.liang@student.rug.nl
Supervisor: Dr. V. Verkhodanova

親愛的參與者您好：

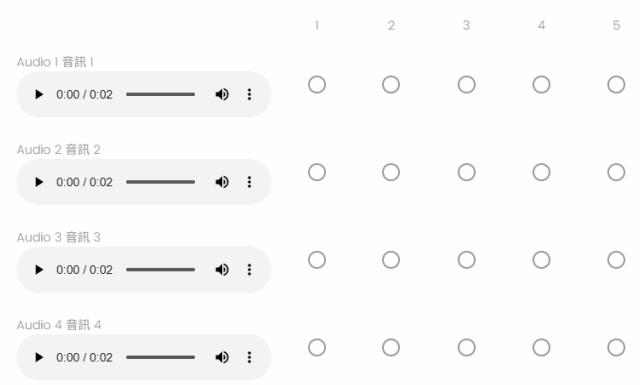
感謝您參與本研究。本研究旨在探討音韻特徵（例如音高 pitch 與能量 energy）在台灣中文與美式英語的語音轉換語音合成系統中的影響，評估不同語音版本在自然度與語調表達上的變化與感知差異。您將會聆聽多組語音樣本，每組包含四個不同版本。每聽完一組語音後，請回答兩個簡短問題。
填答時間：約 5–10 分鐘
參與資格：無需語言學背景或任何專業知識
資料使用：所有資料將匿名記錄，僅供學術研究分析使用，不會收集任何個人識別資訊。

參與完全為自願性，您可隨時中止作答，且不會有任何不利影響。您的資料將依據歐盟《一般資料保護規範（GDPR）》處理。在資料匿名化之前，您有權查詢、更正或刪除您的資料。
如您對本研究有任何疑問，歡迎聯繫
研究生：梁浩維 (Hao-Wei Liang)
荷蘭格羅寧根大學語音科技碩士班
電子郵件：h.w.liang@student.rug.nl

指導教授：Dr. V. Verkhodanova

If you agree to participate, please click “Agree” to begin. If you do not wish to participate, you may simply close this page.
若您願意參與本研究，請點選「同意」開始問卷。若您不同意參與，請關閉此頁面。

Which version of the speech sounds the most natural to you?
以下哪個版本的語音聽起來最自然？
(1=very unnatural, 2=somewhat natural, 3=Neither natural nor unnatural, 4=somewhat natural, 5=very natural)
(1= 非常不自然, 2= 有點不自然, 3= 普通, 4= 稍微自然, 5= 非常自然)



Can you perceive differences in intonation or expressive quality across the different versions?
你能感受到不同版本之間在語調或語音表達上的差異嗎？
(1=No difference at all 完全沒有差異, 5=Very clear difference 差異非常明顯)

[0]https://rug.eu.qualtrics.com/jfe/preview/previewId/8058c298-0198-4709-93f9-b7072843beb4/SV_1Y22P9SOICVUIw6?Q_CHL=preview&Q_SurveyVersionID=current

B Boxplots

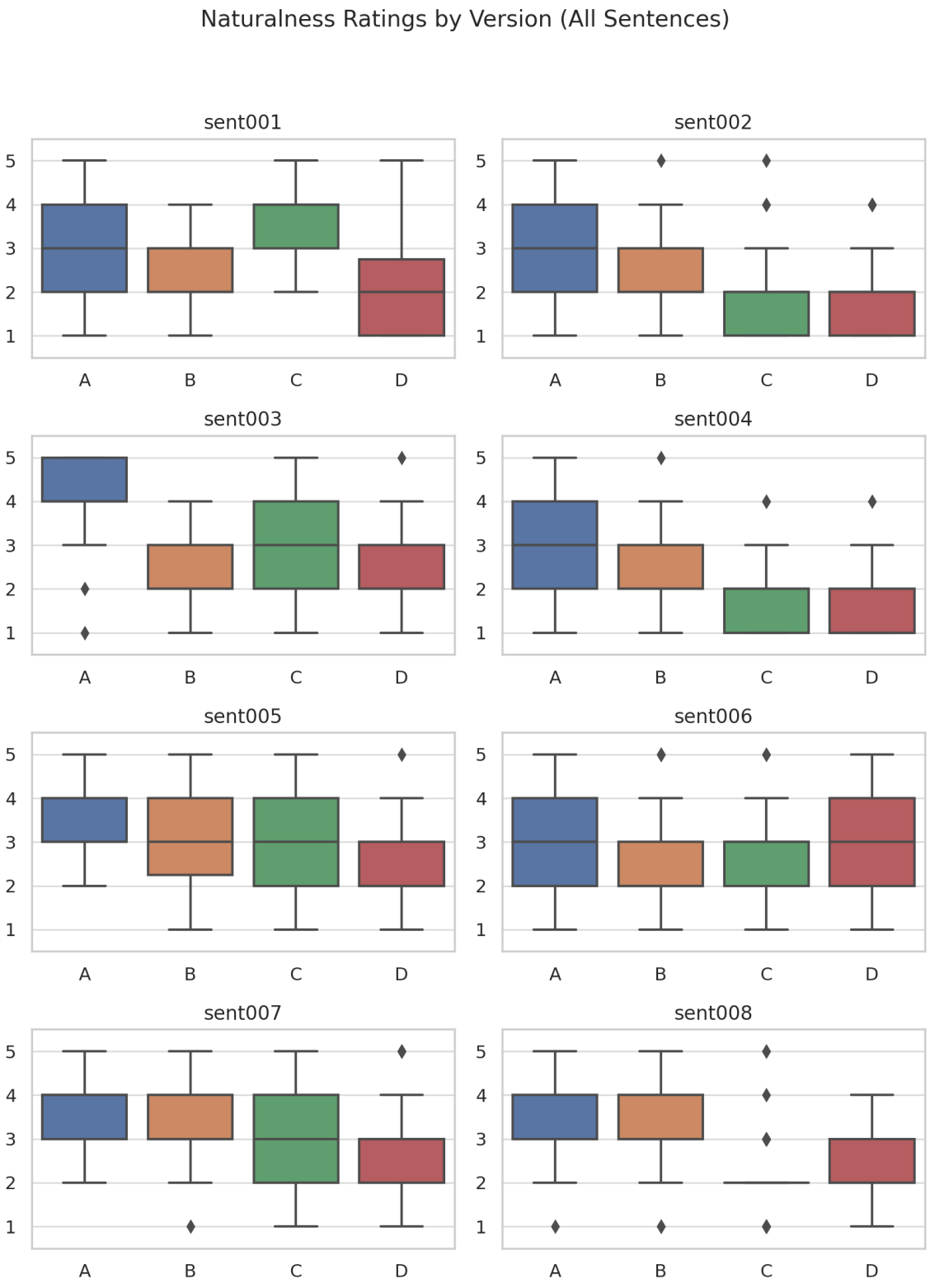


Figure 3: Naturalness Ratings

C AI Declaration

I hereby declare that I am the only author of this Master's thesis and that, to all the best of my knowledge, all work presented herein is my own, except where explicit reference is made to the work of others.

This work has not been submitted for any other degree or professional qualification, unless explicitly stated. All sources of information, including printed materials, online resources, or other forms of work by others, have been properly acknowledged and referenced throughout the thesis.

During the preparation of this thesis, I used ChatGPT (OpenAI, GPT-4, accessed between April and June 2025) for the following purposes:

For Chapters 1 and 2, I used the tool to assist in summarizing and explaining reference studies. The chapters were written by myself, with ChatGPT (OpenAI, GPT-4) support limited to grammar correction and writing refinement.

For Chapter 3, I independently designed the methodology and wrote the full chapter in Mandarin Chinese. I used ChatGPT (OpenAI, GPT-4) for English translation. The final version was fully reviewed and edited by me.

For Section 4.1, I used ChatGPT (OpenAI, GPT-4) to assist with troubleshooting and debugging, and to help describe the hardware of Hábrók. All technical descriptions and content were reviewed and edited by myself.

For Sections 5.2 to 5.4, I used ChatGPT (OpenAI, GPT-4) to generate code for visualizing results using boxplots. The analysis of the results and conclusion were written by myself in both Mandarin Chinese and English and let ChatGPT (OpenAI, GPT-4) translate and check with grammar and spelling.

All content was afterward reviewed, verified, and substantially modified by me.

Date: Jun 11th 2025 Name: Hao-Wei Liang