



university of  
groningen

campus fryslân

# **Enhancing Surprise Perception in TTS through Keyword-Level Prosody Control**

Shuyi Chen



university of  
 groningen

campus fryslân

**University of Groningen - Campus Fryslân**

**Enhancing Surprise Perception in TTS through Keyword-Level Prosody  
Control**

**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  
**Ph.D. Phat Do** (Voice Technology, University of Groningen)

**Shuyi Chen (S5852889)**

June 11, 2025

## Acknowledgements

First and foremost, I'd like to thank my supervisor, Ph.D. Phat Do, for your steady support, thoughtful guidance, and invaluable feedback throughout this thesis. Our weekly catch-ups were especially helpful, as they kept me on track and provided much-needed clarity during both the experimental and writing stages. Your advice at each phase of the project helped shape my thinking and improved the quality of my work.

I'd also like to extend my sincere thanks to all the instructors in the Voice Technology program. Each course offered a unique perspective that helped me build a solid foundation in speech technology, from the technical challenges of synthesis systems to the linguistic principles underlying prosody and emotion. Your passion for teaching and openness to student ideas made the program both intellectually stimulating and genuinely enjoyable. The knowledge and inspiration I gained from your classes were essential in shaping the direction of this thesis.

I am truly grateful to everyone who participated in the listening tests. Your contributions were vital for validating the findings of this project and grounding them in real-world perception. I deeply appreciate your patience in listening to multiple audio samples and providing thoughtful feedback.

Finally, I want to thank my friends and family for being an unwavering source of support throughout this journey. Whether it was helping distribute the survey, assisting with data collection, or simply checking in when things got stressful, your help meant a great deal. I'm especially thankful for your emotional support, constant encouragement, and timely reminders to take care of myself. Thank you for always being there—both in the quiet and the chaotic moments.

## Abstract

Despite major advancements in text-to-speech (TTS) systems, conveying context-sensitive and transient emotions such as surprise remains a persistent challenge. Most existing emotional TTS models rely on global conditioning strategies that apply emotion embeddings uniformly across an utterance. These approaches often fail to capture fine-grained emotional nuances that arise from localized prosodic variation, particularly pitch and energy, which plays a central role in the perception of surprise. This study introduces a lightweight, interpretable prosody control framework that enhances surprise perception by modifying pitch and energy at the keyword level within the FastSpeech 2 architecture. Emotionally salient keywords are automatically identified using a GPT-based semantic detector, and their corresponding pitch and energy values are selectively amplified during inference through the variance adaptor. No model retraining is required, and the method supports configurable emotion types and intensities.

To evaluate the effectiveness and generalizability of the proposed method, a cross-linguistic experimental setup was implemented using both Mandarin Chinese and English. These languages differ in their use of pitch—lexical in Mandarin and intonational in English—offering insights into tonal versus non-tonal prosody control. Synthesized utterances were assessed in a two-part perceptual study involving forced-choice comparisons and scalar ratings of “surprised-ness.”

The results show that prosodic enhancement significantly improves the perception of surprise, with over 90% of listeners preferring the enhanced version in the forced-choice task. English listeners exhibited a linear perceptual response to increasing prosodic intensity, whereas Mandarin listeners showed a non-linear pattern with a perceptual peak at moderate levels. These findings highlight the effectiveness of keyword-level prosody control and underscore the importance of language-specific constraints in emotional speech synthesis.

**Key words:** emotional prosody, text-to-speech, prosody control, surprise perception, cross-linguistic

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Research Questions and Hypotheses . . . . .	8
1.2	Research Contributions . . . . .	8
1.3	Thesis Outline . . . . .	8
<b>2</b>	<b>Literature Review</b>	<b>9</b>
2.1	Background: From Intelligibility to Expressivity in TTS . . . . .	9
2.2	Emotional Speech Synthesis: From Global Labels to Fine-Grained Control . . . . .	10
2.3	Surprise in Speech: Acoustic Correlates and Linguistic Challenges . . . . .	12
2.4	Cross-Linguistic Prosody: Tonal vs. Non-Tonal Challenges . . . . .	13
2.5	Human Perception of Emotion in Synthetic Speech . . . . .	14
2.6	Keyword-Level Prosody Control: A Promising Yet Underexplored Approach . . . . .	15
<b>3</b>	<b>Methodology</b>	<b>16</b>
3.1	Dataset Description . . . . .	16
3.2	Synthesis Models and Prosody Control . . . . .	16
3.3	Technical Implementation Framework . . . . .	17
3.3.1	Smoothing Techniques for Prosodic Consistency . . . . .	18
3.3.2	Structural Modifications to the FastSpeech2 Pipeline . . . . .	19
3.4	Evaluation Methodology . . . . .	19
3.5	Ethics and Research Integrity . . . . .	20
3.5.1	Data Ethics and Privacy . . . . .	20
3.5.2	FAIR Principles Implementation . . . . .	20
3.5.3	Open Science Practices . . . . .	20
3.5.4	Bias and Fairness . . . . .	21
3.5.5	Reproducibility and Replicability . . . . .	21
<b>4</b>	<b>Experimental Setup</b>	<b>22</b>
4.1	Data Preparation . . . . .	22
4.2	Data Splitting . . . . .	22
4.2.1	Development & Test Subsets . . . . .	22
4.2.2	Experiment 1: Perception of Surprise in Mandarin . . . . .	23
4.2.3	Experiment 2: Perception of Surprise in English . . . . .	23
<b>5</b>	<b>Results</b>	<b>24</b>
5.1	Forced-Choice Judgments . . . . .	24
5.2	Open-Ended Comments . . . . .	25
5.3	Rating Scale Analysis . . . . .	25
<b>6</b>	<b>Discussion</b>	<b>28</b>
6.1	Effects of Keyword-Level Prosody Control on Surprise Perception . . . . .	28
6.2	Cross-Linguistic Variation in Prosodic Emotion Perception . . . . .	29
6.2.1	Prosodic Modification Sensitivity and Perceptual Gradient . . . . .	30
6.2.2	Linguistic Constraints and Pitch Functional Load . . . . .	31

---

6.2.3	Emotional Congruence and Keyword Alignment . . . . .	31
6.2.4	Listener Expectations and Cultural Norms . . . . .	32
6.3	Theoretical and Practical Implications . . . . .	33
6.4	Limitations . . . . .	33
6.5	Future Directions . . . . .	34
<b>7</b>	<b>Conclusion</b>	<b>35</b>
	<b>References</b>	<b>36</b>
	<b>Appendices</b>	<b>39</b>
A	Audio Demonstrations . . . . .	39
B	Source Code . . . . .	39
C	Listening Test Sample . . . . .	39
D	AI Usage Declaration . . . . .	40

# 1 Introduction

In recent years, text-to-speech (TTS) technologies have experienced significant advancements, transitioning from traditional concatenative and parametric models to deep learning-based approaches such as WaveNet and FastSpeech2. WaveNet (van den Oord et al., 2018) marked a major breakthrough by synthesizing audio at the waveform level with high naturalness, although its autoregressive nature led to slow inference times. To overcome these limitations, FastSpeech2 emerged as a non-autoregressive, Transformer-based architecture that enables parallel synthesis and introduces explicit variance predictors for prosodic features such as pitch, energy, and duration (Ren et al., 2021).

The controllability of FastSpeech2 has opened new possibilities in expressive TTS, allowing researchers and developers to move beyond flat, neutral speech generation and experiment with emotional or stylistic variation. This shift has significant implications for applications such as audiobooks, virtual assistants, and voice-based storytelling. However, despite promising results in global emotional conditioning, such as synthesizing speech labeled as "angry" or "happy," most systems rely on utterance-level embeddings or style tokens that apply uniformly across the entire sentence. These approaches often fail to capture more nuanced affective states, such as surprise, which are frequently communicated through localized prosodic modulation, especially pitch and energy.

Surprise, as an emotion, is inherently dynamic and context-sensitive. In natural human speech, surprise is often signaled by abrupt pitch rises or shifts concentrated on semantically important keywords. Such fine-grained cues are difficult to reproduce using global control mechanisms alone. Furthermore, surprise is pragmatically complex, varying depending on language structure, discourse context, and listener expectations. To better approximate human-like expressivity, recent studies have proposed localized prosody control. For example, studies have shown (Peters & Almor, 2015) manipulating pitch and energy on sarcasm-related keywords improved listener perception of sarcasm, and the effectiveness of keyword-level pitch enhancement in conveying emotions such as anger and surprise (Diatlova & Shutov, 2023).

Nevertheless, existing work leaves several questions unanswered. First, although there is preliminary evidence supporting keyword-level prosody control, no study has specifically tested whether pitch and energy enhancement can reliably enhance the perception of surprise in synthetic speech. Second, cross-linguistic applicability remains underexplored. This is especially critical for tonal languages such as Mandarin, where pitch carries lexical meaning in addition to prosodic function. A pitch shift intended to signal emotion might conflict with tonal identity, leading to ambiguity or reduced intelligibility.

To address these challenges, this thesis proposes a cross-lingual perception study that tests whether keyword-level pitch and energy manipulation can increase the perceived surprise of synthesized speech. Using FastSpeech 2 as the base model and HiFi-GAN as the vocoder, this study modifies pitch and energy predictor outputs at semantically salient keyword positions during synthesis, generating baseline and pitch- and energy enhanced utterances for evaluation. Listener judgments are collected through a forced-choice task, allowing for controlled comparison of emotional effective-

ness and cross-linguistic generalizability across typologically distinct languages.

## 1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following questions:

**RQ1:** Does keyword-level pitch and energy enhancement increase the perceived surprise of synthetic speech?

**RQ2:** Is the effect of this enhancement consistent across Mandarin and English?

This main question can be broken down into the following hypotheses:

- **H1:** Pitch and energy enhancement on semantically surprise-relevant keywords will significantly increase listeners' perception of "surprise" compared to baseline versions (Diatlova & Shutov, 2023; Yang, Bae, Bak, Kim, & Cho, 2021; Zhou, Xu, & Zhao, 2024).
- **H2:** The improvement in perceived surprise will not differ significantly between Mandarin and English.

## 1.2 Research Contributions

This study introduces a minimalist yet interpretable mechanism for expressing subtle emotional states in multilingual TTS systems. By manipulating pitch and energy at the keyword level instead of across the entire utterance, it offers a finer degree of prosodic control with potentially greater naturalness and listener engagement. Furthermore, it is one of the first studies to compare the perceptual effects of localized pitch and energy adjustment across tonal and non-tonal languages. The results will inform future design of emotionally expressive, language-agnostic TTS systems that require minimal supervision.

## 1.3 Thesis Outline

The structure of this thesis is organized as follows. Section 2 provides the necessary background, including key concepts in expressive speech synthesis and the role of prosody in conveying emotion. It also reviews keyword-level prosody control methods and highlights research gaps in modeling surprise. Section 3 presents the methodology, describing the datasets, synthesis models, pitch and manipulation strategy, and evaluation design. Section 4 outlines the experimental setup, detailing the sentence selection, pitch and energy modification parameters, and listener study procedures for both Mandarin and English. Section 5 presents the results, analyzing listener responses through mean opinion scores and statistical tests. Section 6 discusses these findings in relation to the research questions, examines the implications for cross-linguistic emotional TTS, and identifies future research opportunities. Finally, Section 7 concludes the thesis, summarizing the contributions and their relevance for fine-grained, controllable emotional speech synthesis.



## 2 Literature Review

Recent advances in text-to-speech (TTS) synthesis have enabled the generation of natural-sounding speech across multiple languages and domains. While significant progress has been made in achieving intelligibility and general prosody, the modeling and control of specific emotional expressions—particularly those that are subtle or context-dependent, such as surprise—remain a challenge. This literature review examines key developments in TTS systems with a focus on expressive and emotion-aware synthesis, and explores how linguistic and acoustic features like pitch have been used to model surprise and similar affective states. It also discusses techniques for keyword-level control and the challenges of applying such methods across typologically diverse languages, setting the stage for the proposed research.

This chapter is organized into six subsections. Section 2.1 outlines the transition in TTS research from intelligibility to expressivity, highlighting the role of FastSpeech2 and HiFi-GAN. Section 2.2 reviews emotional speech synthesis, comparing global and fine-grained control strategies. Section 2.3 focuses on the acoustic correlates of surprise and the challenges of modeling it in speech. Section 2.4 examines how prosody is realized differently in tonal and non-tonal languages, with implications for cross-linguistic TTS design. Section 2.5 discusses how human listeners perceive emotional cues in synthetic speech. Finally, Section 2.6 explores keyword-level prosody control as a promising but underexplored method for emotion enhancement in TTS.

### 2.1 Background: From Intelligibility to Expressivity in TTS

The evolution of text-to-speech (TTS) technology has largely been driven by two key objectives: making speech understandable and ensuring it sounds natural. In the early days, systems leaned on concatenative synthesis—essentially stitching together bits of recorded human speech—or used parametric models that generated speech based on statistical patterns. These approaches did manage to make speech clear, but they often came off as robotic or emotionally flat.

TTS development started to shift with the rise of deep learning. Autoregressive models like Tacotron (Shen et al., 2018) and WaveNet (van den Oord et al., 2016) brought in waveform-level synthesis that delivered impressive audio quality, but still suffered from slow inference and offered limited flexibility when it came to control (Ren et al., 2019). This led researchers to look for new approaches. One major response was the development of non-autoregressive models like FastSpeech and its successor, FastSpeech 2. These designs brought faster synthesis thanks to parallel decoding and enabled prosodic features modeling (e.g., pitch, energy, duration.) As a result, the synthesis speed was largely improved, as well as the robustness, without sacrificing quality (Ren et al., 2021).

FastSpeech 2 stands out by enabling more controllable and expressive speech synthesis with integrated prosodic predictors directly into the model. Rather than relying exclusively on text-to-spectrogram transformation, it predicts pitch, energy, and duration separately from the input text, which are then used as conditions in the generation process. Therefore a practical foundation is provided thanks to the separation of prosodic features, and stands out in emotional and stylistic

manipulation, especially when fine-grained expressivity is taken into consideration.

Complementing FastSpeech 2 is HiFi-GAN, a generative adversarial vocoder designed to produce high-fidelity waveforms from mel-spectrograms. HiFi-GAN achieves real-time synthesis and outperforms traditional vocoders in speech quality, making it well-suited for integration into expressive TTS systems (Lim, Jung, & Kim, 2022). Recent systems have explored joint training of FastSpeech 2 and HiFi-GAN, leading to even smoother synthesis pipelines with improved prosody consistency and fewer artifacts.

While these advances significantly improve the naturalness and speaker similarity of synthetic voices, they still struggle with the generation of expressive speech—especially for subtle, context-sensitive emotions like surprise, irony, or disbelief. Expressive speech generation poses unique challenges because emotional cues often manifest as localized prosodic deviations rather than global stylistic shifts. This calls for a shift from sentence-level conditioning (e.g., setting an entire utterance to “happy”) to word-level prosody control, where emotional emphasis is applied selectively to keywords or semantically salient regions.

Current neural architectures, such as FastSpeech 2, are uniquely positioned to support this shift. Their modular control over pitch and energy enables post-synthesis manipulation, allowing researchers to modify acoustic parameters in a localized, interpretable manner—without retraining the full model (Lu, Lee, Wen, Lou, & Oh, 2023). This design makes FastSpeech 2 ideal for research into keyword-level emotion enhancement, forming the technical backbone of the present study.

In summary, modern TTS has evolved from basic intelligibility toward expressive, real-time, and controllable synthesis. Yet, the question of how to systematically manipulate local prosody—particularly pitch—remains underexplored. This work builds upon FastSpeech 2 and HiFi-GAN to investigate whether keyword-level pitch and energy control can effectively enhance the perception of surprise in synthesized speech, across different linguistic contexts.

## **2.2 Emotional Speech Synthesis: From Global Labels to Fine-Grained Control**

The modeling of emotion in text-to-speech (TTS) synthesis has progressed substantially over the past decade. While early attempts focused on conditioning speech output on a single global emotion label (e.g., “happy”, “sad”, “angry”), such methods typically produce coarse-grained emotional expressions that lack contextual nuance. These global approaches often result in exaggerated or homogeneous affect throughout the utterance, failing to reflect how human speakers selectively apply prosodic variation to emotionally salient words or phrases.

In traditional frameworks, emotional control was typically handled by embeddings corresponding to emotion classes. These embeddings were often added to the input sequence or inserted into FastSpeech2’s variance adaptor (Ikeda & Markov, 2024). This method can successfully generate expressive speech, however, because it treats emotion as a broad, sentence-wide effect, its global

nature leads to poor performances when it's applied to subtle emotions such as surprise or irony that tend to surface on specific words rather than across an entire sentence.

Recognizing these limitations, researchers have started leaning toward more detailed, fine-grained emotional control. Diatlova and Shutov (2023) introduced EmoSpeech, a modified FastSpeech2 framework that allows per-phoneme emotional conditioning. Rather than treating emotion as a static, global property, this model dynamically modulates emotion intensity across the utterance, offering greater flexibility and listener-perceived authenticity (Diatlova & Shutov, 2023).

Another key area of focus has been controlling prosody more precisely—especially pitch, energy, and duration. Studies show that emotional content is often communicated most effectively through brief, intense prosodic variations on key lexical items (Mozziconacci, 2002). Take surprise, for example—it's usually conveyed by a sudden jump in pitch on a particular word, not by changing the tone across the whole sentence (Mozziconacci, 2002). This has led to the development of keyword-level control strategies that tweak prosody at the level of individual, emotionally charged keywords, rather than applying blanket changes across the board.

The distinction between global and local prosody modeling is critical. Global prosody refers to sentence-wide features such as overall pitch trend or speaking rate, while local prosody captures dynamic variations at the level of words, syllables, or even vowels (Rao, Koolagudi, & Vempada, 2012). Research has consistently shown that local prosodic features are more effective in distinguishing emotion types, particularly in short utterances or tonal languages.

Lu, Wen, Liu, and Chen (2021) extended this insight to a multi-speaker emotional TTS system that integrates both sentence-level emotion embeddings and fine-grained, disentangled prosody features. Their results confirmed that models incorporating local prosodic variation significantly outperform global-only models in perceived naturalness and expressiveness, without degrading speaker similarity (Lu et al., 2021).

An important advantage of localized prosody control is interpretability. Researchers and system designers can easily understand and manipulate how a change in pitch or energy at a specific word affects emotion perception. This is in contrast to learned global embeddings, which function as black boxes and may interact unpredictably with speaker identity or language.

Despite these advances, few studies have rigorously tested keyword-level pitch and energy control in a cross-linguistic setting. This gap is especially salient given that tonal languages like Mandarin impose lexical constraints on pitch and energy variation, raising questions about how emotional modulation interacts with tone identity.

The present study builds on this line of work by proposing a simple, interpretable approach: manipulating pitch and energy of pre-defined keywords using FastSpeech2, and evaluating its perceptual effect on surprise recognition in both English and Mandarin. This design aims to testify pitch and energy as a cue while maintaining compatibility with existing TTS pipelines.

### 2.3 Surprise in Speech: Acoustic Correlates and Linguistic Challenges

Surprise is a paralinguistic emotion that is often brief, involuntary, and context-dependent. Unlike global emotions such as happiness or sadness that can persist throughout an utterance, surprise typically manifests as a localized prosodic burst, often centered around a key word or phrase. This makes it particularly challenging to model in text-to-speech (TTS) systems, where emotional rendering is typically implemented at the utterance level.

Acoustically, surprise is primarily associated with abrupt rises in pitch ( $F_0$ ), increased pitch range, lengthened syllables, and changes in voice quality, such as breathiness or creakiness. These features often occur on emotionally salient content words. For example, in surprised utterances, the pitch of the initial stressed syllable and the terminal pitch contour are significantly elevated compared to neutral speech. In Estonian, pitch upstepping and extended duration were found to be reliable correlates of surprise, even within syntactically identical interrogative sentences (Asu, Sahkai, & Lippus, 2024).

In Mandarin Chinese, the expression of surprise presents unique challenges due to its tonal nature.  $F_0$  contours are used lexically to distinguish word meanings, which may limit the flexibility available for emotional prosody manipulation. However, research shows that Mandarin speakers can still use pitch excursions beyond tonal norms to convey surprise without confusing lexical identity. X. Liu, Xu, Zhang, and Tian (2021) found that the perception threshold for surprise in Mandarin was around 5 semitones above the neutral baseline, compared to 3 semitones for focus—indicating that surprise occupies a higher prosodic range and is perceptually distinct even within tonal constraints (X. Liu et al., 2021).

Cross-linguistically, the prosodic profile of surprise is remarkably consistent. For instance, studies in Russian and French found that surprised utterances have higher pitch peaks, longer syllable durations, and greater  $F_0$  variance compared to neutral or interrogative speech (Celle & Pélissier, 2022; Makarova, 2000). However, the exact acoustic realizations may vary by language due to constraints imposed by grammar and intonation patterns.

One significant challenge in modeling surprise is disentangling it from other prosodically similar emotions, such as disbelief, emphasis, or sarcasm. All of these can involve pitch prominence and prosodic irregularities. Consequently, the contextual placement and keyword alignment of pitch modulation becomes critical for accurate emotional conveyance. Listeners rely on both acoustic and linguistic cues to interpret surprise, which makes isolated acoustic manipulation (e.g., pitch scaling) a viable but nuanced technique.

In summary, surprise is a locally realized, acoustically complex emotion that interacts strongly with linguistic structure and listener expectation. It can be reliably signaled through pitch and timing, but only when such cues are appropriately aligned with semantic salience. This motivates the approach of the present study: manipulating only the pitch of pre-identified keywords to enhance surprise perception while avoiding overgeneralized or misaligned prosodic patterns.

## 2.4 Cross-Linguistic Prosody: Tonal vs. Non-Tonal Challenges

Prosody—how rhythm, pitch, and stress shape speech—plays a major role in how we express and interpret emotions across different languages. However, the way emotional prosody is realized and the perception of emotional prosody vary substantially between tonal and non-tonal languages due to differences in how pitch and energy is used for linguistic encoding. In non-tonal languages like English, pitch primarily serves an intonational function to indicate things like sentence type, emphasis, or emotional tone. In contrast, in tonal languages such as Mandarin, pitch patterns carry the added responsibility of distinguishing word meanings, creating a unique challenge when pitch also needs to carry emotional weight—for example, in expressing surprise.

Studies have examined how emotional prosody differs across different language types above. Uthiraa and Patil (2023) found that Mandarin speakers tend to use more controlled and predictable pitch patterns—called  $F_0$  contours—due to the tonal structure of the language, whereas English speakers showed much more flexibility in using pitch to express emotion. The finding points out that English speakers seem to lean more on pitch when expressing emotions, while Mandarin speakers make up for pitch constraints by enhancing other prosodic cues such as duration, intensity, or spectral features (Uthiraa & Patil, 2023).

Similarly, Wang, Lee, and Ma (2018) introduced a cross-language production experiment comparing Mandarin and English. It is observed that pitch differences between emotional and neutral utterances were larger in English. On the other hand, Mandarin showed greater modulation in speech rate and phonation cues, such as cepstral peak prominence (CPP) and contact quotient (CQ). These findings support the hypothesis that when one acoustic channel (e.g., pitch) is functionally constrained, speakers naturally shift to using other available signals to get their emotions across (Wang et al., 2018).

From the perceptual side, Mandarin speakers are still capable of identifying emotions from speech, though they may rely on a more distributed acoustic strategy. P. Liu and Pell (2014) found that Mandarin listeners accurately recognized emotional prosody in both their native language and in others (e.g., English, Arabic, German), using universal cues such as  $F_0$  mean and speech rate (P. Liu & Pell, 2014). Yet, the magnitude of  $F_0$  excursions was systematically smaller in Mandarin emotional speech, reinforcing the idea that tonal constraints shape emotional expression and perception.

A recent perceptual study by Xiao and Liu (2024) found that native Mandarin speakers—as well as advanced learners of Chinese—were better at recognizing emotional prosody in Mandarin than monolingual English speakers. However, positive emotions like happiness were less precisely identified, possibly due to overlapping prosodic patterns with lexical tones or subtler acoustic profiles. Moreover, the study also found that emotional recognition improved when the utterances were longer. It seems that having more syllables gives listeners a broader context, helping them tell apart emotional pitch shifts from those tied to word meaning (Xiao & Liu, 2024).

This difference across languages has real implications for how the prosody control is designed. The keyword-level pitch and energy adjustments strategy is applied with caution for tonal vs. non-

tonal languages, in this case, Mandarin and English. In Mandarin, abrupt pitch modifications could inadvertently distort lexical tones, risking intelligibility or unintended meanings. In contrast, English allows greater flexibility for  $F_0$  enhancements without lexical interference. This underscores the need for language-specific thresholds when applying pitch and energy scaling to emotional effects such as surprise.

To sum up, while emotional prosody is based primarily on universal acoustic features such as pitch, energy, and duration, the importance of customizing pitch and energy manipulation strategies to fit the structure of each language should be taken into consideration. These constraints require more targeted, interpretable control strategies when designing cross-linguistic TTS systems capable of conveying emotions without compromising clarity.

## 2.5 Human Perception of Emotion in Synthetic Speech

Human perception of emotion in speech is majorly influenced by prosodic features, including pitch, energy, timing, and intensity. This factor becomes particularly significant in the context of synthetic speech, where machine-generated outputs often lack the nuanced variability found in natural human voices. Nonetheless, empirical evidence indicates that listeners are capable of accurately identifying emotions in synthetic speech when prosodic features are well-regulated and appropriately matched to the intended emotional context.

An early investigation by Kitahara (1988) demonstrated that pitch structure and amplitude contours were more influential than spectral characteristics in conveying emotional content in synthetic speech. The study specifically highlighted the importance of pitch variation in expressing emotions such as joy and anger, while timing cues were found to play a greater role in the perceived intensity of anger.

Further supporting these observations, research by Vlčková-Mejvaldová and Horák (2011) examined the recognition of emotions in Czech synthetic speech. Their findings revealed that listeners relied on a combination of pitch, duration, and intensity to distinguish emotional expressions. Each emotion was shown to exhibit a distinct 'prosodic signature,' and precise manipulation of these acoustic variables was found to significantly affect the accuracy of the recognition.

Perception studies also reveal cross-linguistic similarities and differences. For example, Dimos, Dick, and Dellwo (2015) tested Swiss German and Mandarin Chinese listeners and found that while both groups could recognize basic emotions from prosody, Swiss listeners were more sensitive to subtle gradations of emotion (e.g., different levels of sadness or happiness), suggesting that cultural familiarity with expressive prosody may affect perceptual precision (Dimos et al., 2015).

Interestingly, Ben-David, Multani, Shakuf, Rudzicz, and van Lieshout (2016) showed that prosody often dominates semantics in emotional judgment tasks, especially when the two cues are incongruent. This reinforces the idea that listeners rely heavily on prosody as a primary emotional channel, making it a viable target for manipulation in TTS emotion design (Ben-David et al., 2016).

Together, these findings support the feasibility of prosody-driven emotion synthesis in TTS. They also justify localized manipulation strategies—such as keyword-level pitch and energy control—as perceptually salient methods for enhancing emotional recognition, particularly for momentary and context-dependent states like surprise.

## 2.6 Keyword-Level Prosody Control: A Promising Yet Underexplored Approach

Traditional emotional TTS systems typically apply prosody modifications at the sentence or utterance level, aiming to produce an overall emotional tone (e.g., “happy” or “angry” speech). However, such global approaches often fail to capture the fine-grained emotional fluctuations found in natural speech, particularly in conveying emotions like surprise, which tend to be locally concentrated on specific words or phrases.

Recognizing this, recent advances have moved toward word-level and even phoneme-level control of prosodic parameters such as pitch ( $F_0$ ), energy, and duration. For example, Guo, Du, and Yu (2022) proposed an unsupervised word-level prosody tagging framework that significantly improved speech expressiveness and controllability over vanilla FastSpeech2. Their work demonstrated that explicitly labeling words with prosodic intent enables more flexible, interpretable emotional synthesis, especially during inference-time manipulation (Guo et al., 2022).

Similarly, Lee and Kim (2019) introduced temporal prosody embeddings into end-to-end TTS networks, allowing users to manipulate pitch and amplitude at both frame and phoneme levels. Their approach empowered frame-level prosody control, but required careful supervision or complex network tuning to align prosodic patterns with semantics (Lee & Kim, 2019).

The keyword-level approach offers a practical compromise between flexibility and interpretability. By selectively modifying pitch and energy on pre-identified emotionally salient keywords, researchers can test the perceptual effects of prosody in a focused, reproducible manner. This method has been shown to produce significant improvements in perceived emotional intensity with minimal changes to overall speech rhythm or semantics (Luo, Takamichi, Saito, Koriyama, & Saruwatari, 2024).

Furthermore, few works have tested such localized control strategies in cross-linguistic contexts. Tonal languages like Mandarin pose unique challenges for prosodic manipulation due to the lexical role of  $F_0$ . Yet, no studies to date have systematically compared the effectiveness of keyword-level pitch and energy control across tonal and non-tonal languages.

This gap forms the basis of the present study: to investigate whether pitch and energy enhancement on semantically crucial keywords can increase perceived surprise in TTS-generated speech across Mandarin and English. The approach provides both interpretability for analysis and flexibility for future integration into end-to-end synthesis pipelines.

### 3 Methodology

This chapter presents the methodology adopted in this study to investigate whether keyword-level pitch and energy control can enhance the perception of surprise in synthetic speech. The methodological approach is organized into five sections: (3.1) dataset description, (3.2) synthesis models and prosody control techniques, (3.3) technical implementation framework, (3.4) evaluation procedures, and (3.5) ethical considerations and open science practices.

#### 3.1 Dataset Description

To support both Mandarin and English speech synthesis, four publicly available speech corpora were used. For Mandarin, the DataBaker Mandarin Speech Corpus and AISHELL-3 were selected. The DataBaker corpus is a single-speaker dataset containing approximately 12 hours of professionally recorded utterances, totaling around 10,000 sentences. In contrast, AISHELL-3 is a large-scale multi-speaker corpus with over 85 hours of speech from more than 200 speakers, offering greater variability and speaker diversity.

For English, two corpora were used: LJSpeech and LibriTTS. LJSpeech is a single-speaker dataset consisting of over 13,000 audiobook utterances, totaling approximately 24 hours of speech. LibriTTS, by contrast, is a multi-speaker corpus comprising over 100 hours of speech derived from public domain audiobooks, offering a broad range of voices and speaking styles.

All audio files were resampled to 16kHz, and phoneme-level alignments were generated using the Montreal Forced Aligner (MFA). Each dataset was used to train an independent FastSpeech2 model tailored for language-specific synthesis.

#### 3.2 Synthesis Models and Prosody Control

FastSpeech2 was selected as the core synthesis model due to its non-autoregressive architecture and integrated predictors for pitch, energy, and duration. HiFi-GAN was adopted as the vocoder to convert mel-spectrograms into high-quality waveforms. To achieve surprise-specific emotional expressiveness, a keyword-level prosody control mechanism was developed.

The prosody control framework consists of three primary components. First, a keyword detection function based on the GPT-4o API was implemented to identify up to three emotionally salient words related to surprise. Second, these keywords were aligned to phoneme indices via forced alignment. Third, during inference, pitch and energy values at the identified keyword regions were scaled according to the specified emotion level.

Instead of fixed scaling, the control values were dynamically set using command-line arguments (`–emotion type` and `–emotion level`). Emotion intensity was mapped to predefined multipliers: mild



(1.2×), moderate (1.3×), and strong (1.5×). These adjustments were applied at the phoneme level through the variance adaptor module in FastSpeech2.

The choice of the 1.3× scaling factor for pitch and energy was informed by prior perceptual and acoustic studies on the expression of “surprise” in speech. Research in Mandarin Chinese has shown that the perception of surprise requires a pitch increase of at least 5 to 7 semitones from the neutral baseline. This corresponds roughly to a linear frequency multiplier in the range of 1.3 to 1.4, depending on the base pitch (X. Liu & Xu, 2016; X. Liu et al., 2021). These studies demonstrate that listeners begin to reliably perceive the emotion of surprise only after this threshold is crossed, suggesting that a 1.3× increase is a perceptually grounded and conservative estimate for moderate emotional intensity.

Moreover, increasing pitch alone can sometimes result in unnatural-sounding synthesis. To address this, we adopted a joint modification approach that scales both pitch and energy simultaneously, a method supported by findings in expressive TTS research. Specifically, Sorin, Shechtman, and Pollet (2015) showed that simultaneous adjustment of pitch and energy preserves the natural prosodic balance and avoids the common artifacts introduced by independent pitch manipulation (Sorin et al., 2015). This approach enhances emotional clarity while maintaining the quality and coherence of synthesized speech.

In summary, the prosody control mechanism’s dynamic scaling design—particularly the moderate setting of 1.3×—is grounded in both perceptual thresholds and best practices in emotional speech synthesis. This ensures that the expression of “surprise” is both perceptually effective and acoustically natural.

### 3.3 Technical Implementation Framework

The synthesis pipeline was structured into three stages. In Stage 1, FastSpeech2 and HiFi-GAN models were trained using phoneme-segmented text and corresponding audio. In Stage 2, the GPT-based keyword detector identified surprise-related words, which were then mapped to phoneme positions using forced alignment. In Stage 3, pitch and energy vectors at keyword positions were adjusted during inference.

These three stages—model training, keyword detection and alignment, and inference-time prosody modulation—are integrated into a unified synthesis framework. The full system pipeline is shown in Figure 1, highlighting the flow of inputs and the locations of key modifications to the standard FastSpeech 2 architecture.

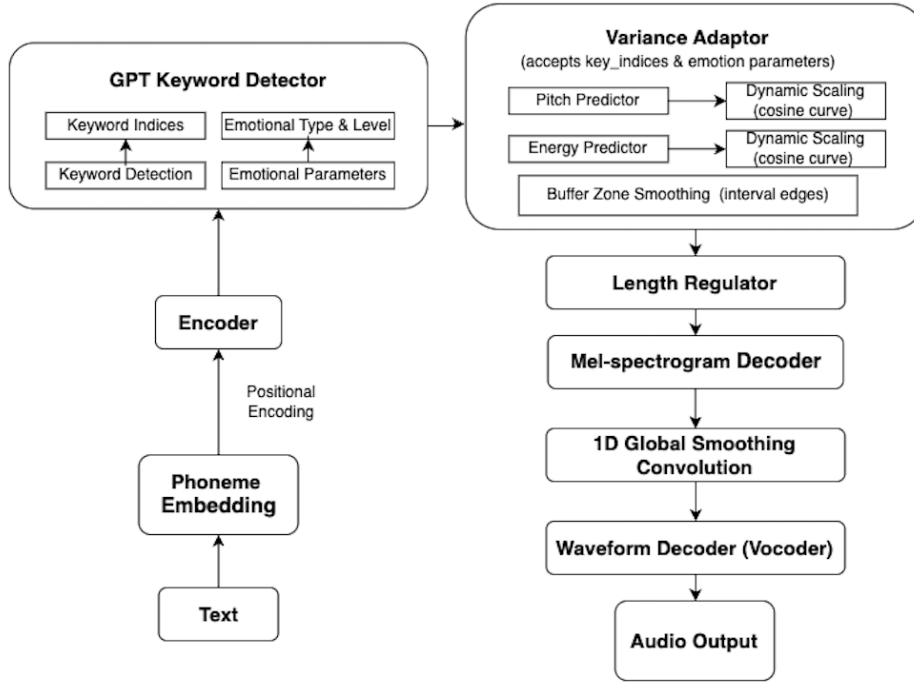


Figure 1: The proposed keyword-level prosody control pipeline

### 3.3.1 Smoothing Techniques for Prosodic Consistency

To enhance the perceptual continuity of prosodic transitions—particularly around emotionally emphasized keyword regions—two post-processing smoothing techniques were introduced. These techniques aim to address both local discontinuities and global inconsistencies that may arise from abrupt pitch or energy modulation.

The first technique focuses on local transitions at the boundaries of prosodically enhanced keywords. Abrupt shifts in pitch or energy across neighboring frames can result in unnatural artifacts, especially when the emotional intensity is high. To mitigate this, a lightweight transition smoothing mechanism was applied: pitch and energy contours are adjusted by interpolating values across a small window before and after the keyword segment, ensuring a gradual onset and offset of emphasis. The window length and interpolation weights are heuristically set to balance naturalness and control.

The second technique targets global consistency by smoothing the entire mel-spectrogram along the temporal axis. Specifically, a one-dimensional convolutional kernel is applied to each frequency band across time, effectively attenuating high-frequency fluctuations caused by localized modulations. This process improves the overall spectral cohesion of the utterance without degrading emotional expressiveness.

These two smoothing strategies were applied sequentially prior to waveform synthesis. The first

reduces local acoustic discontinuities, while the second ensures global spectral smoothness. A high-level illustration of this two-stage smoothing pipeline is shown in Figure 1. For implementation-level details and parameter configurations, please referred to the open-source repository<sup>1</sup>.

### 3.3.2 Structural Modifications to the FastSpeech2 Pipeline

To support flexible and user-defined emotional customization, several architectural extensions were introduced into the FastSpeech2 synthesis pipeline.

First, a gpt-based keyword detection module was added to identify emotionally salient words based on the specified emotion type. This module queries the OpenAI GPT API using custom prompts to return up to three keywords per sentence. If fewer than three are found, it fills the remaining slots with 'None'. The keywords are then mapped to their respective phoneme spans for alignment.

Second, the emotion type argument allows users to specify the target emotion category (e.g., surprise, joy, anger). This enables future expansion to other emotions and promotes modularity in prosody control.

Third, emotional intensity control is realized via the emotion level argument. The system maps each intensity level to predefined prosody scaling values using a dictionary. For instance, 'mild' corresponds to pitch  $\times 1.2$  and energy  $\times 1.2$ , 'moderate' to  $\times 1.3$ , and 'strong' to  $\times 1.5$ . This parameterized approach standardizes intensity modulation and avoids manual intervention.

Finally, these features were fully integrated into the inference routine, where detected keywords trigger prosody modifications, followed by smoothing steps, and finally waveform synthesis. This unified architecture allows end-to-end generation of emotionally expressive speech with controllable and reproducible prosodic effects.

## 3.4 Evaluation Methodology

In the first perceptual task, participants listened to paired utterances labeled Audio A and Audio B. One version was generated using the baseline FastSpeech 2 model, while the other incorporated keyword-level pitch and energy enhancement. The assignment of A and B was randomized across participants. Listeners were asked to indicate which version sounded more surprised, or select "unsure" if no clear difference was perceived. An optional comment box was provided for participants to explain what influenced their judgment.

The second task asked participants to rate four versions of the same sentence—baseline, mild, moderate, and strong—on a 1–10 scale of perceived surprise. The presentation order of the four audio clips within each set was randomized, and participants rated each clip independently.

---

<sup>1</sup><https://github.com/S-CHEN-rug/Surprise-Perception>

Altogether, the perception test consisted of 20 sentence pairs in Task 1 (10 Mandarin and 10 English), and 20 sentence sets in Task 2 (10 Mandarin and 10 English), each containing four prosodically varied versions.

Forced-choice responses were analyzed using one-sample binomial tests to determine whether enhanced versions were selected significantly more often than chance (50%). No human speech was included in the stimuli; all utterances were synthesized using the FastSpeech 2 system with HiFi-GAN vocoding.

### **3.5 Ethics and Research Integrity**

This study was conducted in accordance with institutional ethical guidelines, all experiments involving human participants followed informed consent procedures, with anonymity preserved through coded identifiers. Data collected through the perception tests contained no personally identifiable information and were stored securely on university-managed servers. All datasets used for synthesis were publicly available and licensed for academic use.

#### **3.5.1 Data Ethics and Privacy**

All data used in this study complies with data protection regulations and institutional ethical standards. The speech datasets selected for synthesis—DataBaker, AISHELL-3, LJSpeech and LbriTTS—are publicly released under permissive licenses and contain no personally identifiable information. For the perception study, participants were recruited voluntarily and provided informed consent prior to the experiment. No demographic or biometric data were collected. Data access was limited to authorized researchers, and all recorded responses were anonymized and stored on password-protected servers. No third-party data with restricted access or proprietary constraints were utilized.

#### **3.5.2 FAIR Principles Implementation**

The research process adheres to the FAIR principles to ensure that data and resources are Findable, Accessible, Interoperable, and Reusable. All code and documentation are shared via a public repository with clear versioning and licensing. Reusability is supported through detailed documentation and guidelines for future users.

#### **3.5.3 Open Science Practices**

In alignment with open science practices, all source code, experiment configurations, and analysis notebooks will be made available on a public GitHub repository under an open-source MIT license. The repository includes detailed README files, usage instructions, and environment specifications. Synthesized audio samples and perception test materials will be shared, along with links in the appendix. Version control is managed using Git, and contributions are tracked to support transparency.

and collaborative development. All resources will remain publicly accessible after the completion of the project.

#### **3.5.4 Bias and Fairness**

Potential biases in both data and model output are recognized and mitigated to the extent possible. The selected speech datasets represent both male and female speakers in both languages, which could avoid gender-related acoustic patterns. No emotion labels were present in the training data, reducing risks of label-induced bias. However, cultural variation in emotion perception—particularly surprise—may influence listener judgments. This limitation is addressed by balancing participant demographics across language backgrounds and reporting analysis disaggregated by language group. Future extensions may consider age and dialect diversity to further assess algorithmic fairness.

#### **3.5.5 Reproducibility and Replicability**

Reproducibility and replicability were prioritized in the design and implementation of this study. All source code, training configurations, and evaluation tools are hosted in a public GitHub repository. The repository includes detailed instructions for replicating the entire synthesis and evaluation pipeline, from data preparation to result analysis.

## 4 Experimental Setup

To ensure full reproducibility of this research, the experimental setup is documented in detail. This section provides an overview of data preparation, data splitting strategies, and the configuration of two controlled experiments involving pitch and energy-enhanced TTS synthesis. All procedures are accompanied by exact parameters, software versions, and implementation choices. Source code and configuration files used in the experiments are available in a public GitHub repository, which includes automation scripts, environment specifications, and synthesized audio outputs.

### 4.1 Data Preparation

The Mandarin Chinese single-speaker data used in this study was extracted from the DataBaker corpus, consisting of 10,000 single-speaker recordings in WAV format at 16,000 Hz. The English single-speaker data was selected from the LJSpeech corpus, recorded by a US English speaker in the same format. Pre-trained AISHELL-3 (Mandarin Chinese multiple-speaker) and Libri-TTS (English multiple-speaker) checkpoints are also used for speech synthesis to avoid gender bias. The grapheme-to-phoneme (G2P) conversion are manually checked with the provided transcription, and punctuation was removed. Phoneme alignment was performed using Montreal Forced Aligner (MFA) with corresponding pretrained dictionary, acoustic model, and G2P model.

For pitch and energy manipulation, each target sentence contained two to three semantically salient keyword eligible for pitch and energy enhancement. Pitch and energy contours of the keywords were modified using FastSpeech2 pitch and energy control scaler, applied only to the keyword region according to the corresponding emotion level setting.

### 4.2 Data Splitting

Each dataset was split into 80% training, 10% validation, and 10% test sets at the utterance level to ensure phonetic and speaker consistency. The test set was reserved for perceptual stimuli generation. Random seeds for shuffling and splitting were fixed at 42 for reproducibility.

#### 4.2.1 Development & Test Subsets

A total of 10 sentences were selected from the test sets (5 Mandarin, 5 English). Selection criteria included (1) moderate length (6–10 syllables), (2) neutral syntactic structure, and (3) presence of a semantically identifiable keyword associated with surprise (e.g., “really” / “jīng rán”). Keywords were selected based on part-of-speech tagging and semantic salience, and manually verified for clarity. All selected sentences were synthesized in four versions: (1) baseline (no modification), and (2) pitch and energy enhanced (emotion level mild, moderate, or strong).

### 4.2.2 Experiment 1: Perception of Surprise in Mandarin

This experiment investigated whether keyword-level pitch and energy enhancement increases the perceived surprise of synthetic Mandarin utterances. Ten sentences were selected and synthesized using FastSpeech 2 and HiFi-GAN. Pitch and energy enhancement was applied via the FastSpeech 2 scaler module, with structural adjustments to amplify prosodic features at the keyword region relative to the original  $F_0$ . Each sentence was rendered under four prosodic conditions: baseline, and enhanced emotion levels (mild, moderate, and strong).

Two perceptual evaluation tasks were conducted. In the forced-choice task, native Mandarin-speaking participants were presented with 10 pairs of utterances in randomized order, each pair consisting of a baseline and one enhanced version. Participants were instructed to select the version that sounded more surprised. Responses were analyzed using the binomial test to assess preference significance.

In the scalar rating task, participants rated all four versions of each sentence on a 10-point scale of perceived surprise. These scores were used to evaluate sensitivity to prosodic manipulation and to explore the relationship between enhancement intensity and emotional perception. Visual inspection of rating patterns revealed a non-linear trend in Mandarin: moderate enhancements received the highest scores, while strong enhancements were sometimes rated lower than mild ones.

A total of 73 native Mandarin-speaking participants completed both tasks via an online survey platform. Their responses formed the basis for subsequent quantitative and qualitative analyses.

### 4.2.3 Experiment 2: Perception of Surprise in English

This experiment mirrored the Mandarin setup using 10 English sentences drawn from LJSpeech and a subset of LibriTTS. Each sentence was synthesized under the same four prosodic conditions. Keyword-level pitch and energy enhancement was applied using the same FastSpeech 2 modification strategy.

Fifty-seven native English-speaking participants completed both the forced-choice and scalar rating tasks through an online interface. In the forced-choice task, each pair consisted of a baseline and one enhanced version, and participants were asked to select the one that sounded more surprised. In the scalar rating task, participants rated all four versions of each sentence—baseline, mild, moderate, and strong—on a 1–10 scale of perceived surprise.

The results from both language conditions were compared to evaluate the effectiveness of prosodic enhancement strategies and to assess cross-linguistic generalizability between tonal (Mandarin) and non-tonal (English) prosodic systems.

## 5 Results

This chapter outlines the planned analysis procedures and the expected outcome structure of the perception experiment. The results will be presented in three sections: (5.1) participants' forced-choice judgments regarding which utterance sounded more surprised; (5.2) qualitative insights derived from optional open-ended comments; and (5.3) scalar ratings of perceived surprise across different levels of prosodic enhancement. These analyses aim to evaluate whether keyword-level pitch and energy manipulation successfully enhances the perception of surprise, and whether such effects are consistent across language groups (Mandarin and English).

### 5.1 Forced-Choice Judgments

Participants completed a forced-choice task in which they listened to two versions of the same sentence: one baseline utterance and one prosodically enhanced version with increased pitch and energy on emotionally salient keywords. For each pair, participants selected the version that sounded more surprised or indicated uncertainty if they could not decide.

The results show a clear preference for the prosodically enhanced version across both English and Mandarin listener groups, with selection rates consistently above 85%, well exceeding the 50% chance level. This suggests that the proposed keyword-level enhancement strategy is perceptually effective in conveying surprise. One-sample binomial tests confirmed that the enhancement condition was significantly preferred. While English listeners showed slightly more consistent preference patterns, Mandarin listeners also exhibited strong responses, especially for sentences where pitch and energy enhancement aligned with natural tone contours.

Table 1: Forced-choice Accuracy per sentence

<b>Mandarin</b>					
<b>Female</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
<b>Accuracy</b>	97.26%	91.78%	95.89%	97.26%	97.26%
<b>Male</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
<b>Accuracy</b>	97.26%	95.89%	96.49%	97.26%	98.63%
<b>English</b>					
<b>Female</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
<b>Accuracy</b>	98.25%	92.98%	91.23%	85.96%	92.98%
<b>Male</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
<b>Accuracy</b>	92.98%	94.74%	91.23%	98.25%	85.96%



## 5.2 Open-Ended Comments

Following each forced-choice trial, participants were optionally invited to provide a brief explanation of their decision. Thematic analysis of these comments reveals several recurring perceptual themes that offer insight into how listeners interpret prosodic enhancements.

English participants frequently used descriptors such as "clicky," "robotic but higher," or "slightly stressed" to characterize the modified utterances. While some noted that the difference was "not so obvious" or showed "no big difference," others acknowledged perceivable changes in pitch or vocal emphasis that signaled emotional salience.

Mandarin participants also reported subtle yet noticeable changes, using phrases like "the intonation felt somewhat strange but had more variation", "the pitch rose in the middle", and "there was a rise in pitch". Some responses, such as "I was just guessing", reflected cases where the perceptual distinction was less pronounced, but the trend still leaned toward identifying the prosodically enhanced version.

Notably, some participants directly referenced the manipulated keywords, indicating that they could identify which words had been enhanced, which aligns with the actual keywords whose pitch and energy were enhanced. This suggests that the keyword-level manipulation was salient enough to be consciously detected by listeners, further validating the interpretability of the proposed control strategy.

Table 2: Subjective Open-Ended Comments (except enhanced keywords)

Subjective Open-Ended Comments	
Mandarin	English
The tone is a bit weird but can tell the difference	Sound a bit clicky
Not so obvious	Robotic but higher?
I took a guess	Slightly stress
Seems to be no big difference	Not so obvious
A bit higher in the middle	No big difference
Higher pitch	

## 5.3 Rating Scale Analysis

In addition to the forced-choice task, participants also completed a scalar rating task in which they evaluated four versions of the same sentence—namely, a baseline utterance and three prosodically enhanced variants with increasing levels of pitch and energy modulation, labelled as Mild, Moderate,

and Strong. Each version was rated on a 0 to 10 scale, where 0 indicated “not surprised at all” and 10 indicated “extremely surprised.”

To control for potential speaker-related bias, both male and female voices were included in the stimuli. As shown in Table X, the rating results reveal a clear upward trend in perceived surprise from the Base to the Moderate condition, followed in some cases by a decline at the Strong level. This general pattern holds across most sentences and speaker genders, but with notable cross-linguistic differences.

For English stimuli, both male and female voices exhibited consistent increases in surprise ratings as prosodic intensity increased. The Strong condition achieved the highest scores in nearly all English sentences. For instance, English Female Sentence 3 increased from 3.96 (Base) to 8.21 (Strong), and English Male Sentence 10 rose from 4.02 (Base) to 7.84 (Strong). The progression from Base → Mild → Moderate → Strong is largely monotonic, suggesting that in English, stronger prosodic cues reliably enhance the perception of surprise. No noticeable degradation or perceptual confusion was observed even at the highest intensity level. This supports prior findings that English, as a non-tonal language, is more tolerant of expressive pitch and energy modulation, especially in emotional speech synthesis.

In contrast, Mandarin ratings displayed a different pattern. While Base-to-Moderate enhancements reliably increased perceived surprise, the Strong condition often did not lead to further improvement and, in several cases, resulted in a drop in ratings. For example, Mandarin Female Sentence 2 peaked at the Moderate level (6.82), then dropped in the Strong condition (4.08), and Mandarin Male Sentence 6 was rated highest in the Moderate condition (7.75), with a lower score for Strong (6.74). This decline may be attributed to the interference between aggressive pitch and energy manipulation and lexical tone recognition in Mandarin, where pitch variation plays a dual role in signaling both prosody and lexical meaning. Excessive modification may have introduced perceptual ambiguity or unnaturalness that reduced emotional salience.

Interestingly, across both languages, gender effects were minimal. That is, ratings from female- and male-synthesized voices followed similar trends, and no systematic difference in magnitude or direction was observed between genders. The inclusion of both genders served primarily to ensure generalizability and reduce voice-specific bias rather than to test for gender-based differences. The primary variation arose not from gender, but from language type—with English listeners responding favorably to increased prosodic enhancement, and Mandarin listeners showing a preference for moderate levels, beyond which the effectiveness of prosodic cues plateaued or declined.

In summary, scalar ratings support the hypothesis that prosodic enhancements increase perceived emotional intensity. However, the optimal level of enhancement is language-dependent: English benefits from stronger cues, while Mandarin requires more conservative modulation due to tonal constraints.

Table 3: Scalar Ratings by Condition, Language, Gender, and Sentence

<b>Mandarin</b>					
<b>Female</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
Strong	3.75	4.08	2.32	6.71	3.59
Moderate	7.18	6.82	6.48	5.10	5.15
Mild	5.44	5.37	4.82	3.86	2.79
Base	2.49	3.10	3.64	2.53	6.21
<b>Male</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
Strong	3.71	2.56	3.81	6.67	3.58
Moderate	6.78	6.68	6.67	5.18	6.79
Mild	5.22	5.15	5.08	3.89	4.82
Base	2.44	3.99	2.62	2.66	2.22
<b>English</b>					
<b>Female</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
Strong	8.00	7.96	8.21	6.49	6.68
Moderate	6.53	6.93	6.89	7.79	8.09
Mild	5.23	5.44	5.82	5.49	5.60
Base	4.09	4.19	3.96	4.19	4.25
<b>Male</b>	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
Strong	6.74	7.65	7.98	7.75	7.84
Moderate	7.75	6.74	6.68	6.61	5.47
Mild	4.19	5.30	5.33	4.98	6.30
Base	5.44	4.04	3.93	4.26	4.02

## 6 Discussion

This chapter interprets the experimental findings in light of the study’s core research questions and hypotheses. The discussion integrates empirical outcomes with theoretical frameworks in expressive speech synthesis and cross-linguistic prosody, offering insights into the design of interpretable, emotion-aware TTS systems. Each subsection addresses a distinct dimension of the research, including perceptual outcomes, linguistic constraints, system-level considerations, and directions for further investigation.

### 6.1 Effects of Keyword-Level Prosody Control on Surprise Perception

The results from both the forced-choice and scalar rating tasks provide strong empirical support for Hypothesis 1 (H1): keyword-level pitch and energy enhancement can significantly increase the perception of surprise in synthetic speech. In the forced-choice task, enhanced utterances were overwhelmingly preferred over baseline counterparts, with selection rates exceeding 90% across both English and Mandarin sentences. This indicates that even localized prosodic modifications—when aligned with semantically salient keywords—are sufficient to produce perceivable emotional effects. This finding extends the work of Diatlova and Shutov (2023), who demonstrated that fine-grained prosody control at the phoneme level can enhance perceived emotional intensity in synthetic speech, and supports the argument by Mozziconacci (2002) that emotions like surprise are best conveyed through localized, rather than global, prosodic cues.

The scalar rating task further revealed that listeners are sensitive to gradations in prosodic intensity. In English, ratings increased monotonically across the Base → Mild → Moderate → Strong conditions, indicating a consistent perceptual mapping between the degree of prosodic enhancement and perceived emotional intensity. This linear trend aligns with previous findings in non-tonal language contexts, where pitch excursions are commonly interpreted as expressive cues without affecting intelligibility (Lu et al., 2021; Mozziconacci, 2002). While Mandarin showed less consistent trends, moderate enhancement still produced significantly higher surprise ratings than the Base or Mild conditions in most cases, confirming the perceptual efficacy of controlled pitch and energy modulation even in tonal contexts. This partially supports the findings of X. Liu et al. (2021), who reported that Mandarin listeners can detect surprise through pitch modifications, though within narrower acoustic bounds due to tonal constraints.

These findings validate the central premise of the study: targeted, interpretable prosodic manipulation can enhance emotional expressivity in TTS without altering model architecture or requiring retraining, especially when such modifications are grounded in linguistic salience. Moreover, the successful application of this technique across both tonal and non-tonal languages suggests a promising direction for designing flexible, language-aware emotional TTS systems, echoing calls from Guo et al. (2022) and Luo et al. (2024) for more controllable and interpretable prosody frameworks in expressive speech synthesis.

## 6.2 Cross-Linguistic Variation in Prosodic Emotion Perception

Although both language groups benefited from prosody enhancement, systematic differences were observed in how pitch and energy manipulation affected surprise perception—partially confirming Hypothesis 2 (H2). These findings contribute to a growing body of cross-linguistic research on emotional prosody perception, and particularly support the notion that the functional load of pitch in a language strongly modulates how prosodic cues are interpreted (Uthiraa & Patil, 2023; Wang et al., 2018).

In English, where pitch plays an intonational rather than lexical role, listeners responded positively to increasing prosodic intensity. Strong enhancement levels yielded the highest scalar ratings, suggesting that expressive pitch contours aligned with emotionally marked keywords are not only perceptually salient but also cognitively congruent with the prosodic expectations of English. This observation is consistent with prior studies demonstrating that English speakers are more tolerant of exaggerated pitch excursions for emotional expression, as pitch in English primarily functions to signal discourse-level information such as emotion or emphasis (Ben-David et al., 2016; Mozziconacci, 2002).

In Mandarin, by contrast, the relationship between enhancement level and perceived surprise was non-linear. While moderate enhancement generally improved emotional perception, strong enhancements sometimes resulted in decreased ratings—a pattern not observed in English. This result aligns with findings by X. Liu et al. (2021), who reported that Mandarin listeners could detect surprise only when pitch exceeded a 5–7 semitone threshold, but that excessive pitch variation risks conflicting with lexical tone identity. The trade-off observed here reinforces the dual role of pitch in Mandarin as both a lexical and affective signal, and supports claims by P. Liu and Pell (2014) that tonal language speakers rely more on a distributed set of acoustic cues to interpret emotional content.

As shown in Figure 2, scalar ratings in English demonstrated a monotonic increase across emotion levels, with little overlap between box quartiles, suggesting consistent perceptual gain from prosodic enhancement. In Mandarin, however, the boxplots revealed an inverted-U pattern: while mild and moderate enhancements raised surprise perception, strong modifications often resulted in a drop, with wider interquartile ranges and visible outliers. This supports the hypothesis of a perceptual ceiling in tonal languages (Xiao & Liu, 2024), where aggressive pitch scaling may exceed the phonological tolerance for  $F_0$  modulation and disrupt lexical interpretation.

Notably, the lack of overlap between interquartile ranges in English indicates a stronger perceptual separation between prosodic conditions. In contrast, Mandarin responses showed greater variability, suggesting individual differences in tone processing tolerance, a finding that mirrors previous observations of cross-speaker variability in tonal emotion perception (Wang et al., 2018). These observations further align with cultural norm theories, which suggest that Mandarin-speaking listeners may exhibit more conservative expectations for expressive intonation, especially when prosodic contours deviate sharply from native tonal patterns (Dimos et al., 2015; Uthiraa & Patil, 2023).

These cross-linguistic differences underscore the importance of designing TTS systems that are not

only prosodically controllable, but also language-aware, adapting to the structural and perceptual constraints of the target language. Emotionally expressive synthesis in tonal languages like Mandarin must balance affective salience with tonal integrity, requiring more nuanced parameterization than in non-tonal contexts.

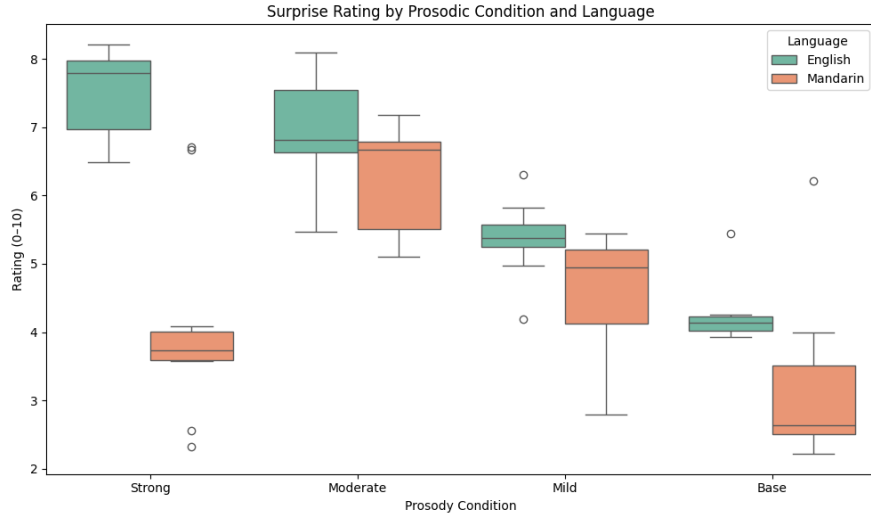


Figure 2: Surprise Rating by Prosodic Condition and Language

### 6.2.1 Prosodic Modification Sensitivity and Perceptual Gradient

In English, listeners exhibited a clear linear and monotonic perceptual response to increasing prosodic intensity. As shown in Figure 2, scalar ratings rose consistently from the Base to the Mild condition, then to Moderate, and peaked under the Strong condition. This progressive increase suggests that prosodic cues such as pitch and energy are readily interpreted as emotional intensifiers in English, where pitch serves primarily intonational and pragmatic functions. These findings align with Mozziconacci (2002), who emphasized that pitch elevation on emotionally salient segments enhances affective interpretation without compromising intelligibility in non-tonal languages. The compressed spread and elevated median ratings for Strong prosody further indicate high listener consensus and tolerance toward exaggerated  $F_0$  contours—a pattern similarly reported by Ben-David et al. (2016), who found that prosody often dominates semantics in emotional judgments when the cues are sufficiently strong and well-aligned with discourse context.

Such pitch excursions in English are likely perceived as congruent with the emotional function of specific keywords (e.g., “really?”), reinforcing the expression of surprise without introducing semantic ambiguity or unnaturalness. This also reflects the flexibility of pitch in English as a resource for emotional emphasis, corroborating Lu et al. (2021)’s conclusion that localized pitch variations outperform global emotional conditioning in producing expressive synthetic speech.

In contrast, the Mandarin data reveal a non-linear pattern, resembling an inverted-U curve. While Mild and especially Moderate enhancements elicited perceptually stronger surprise ratings, the

Strong condition resulted in a noticeable drop in median scores and increased dispersion, as illustrated in the right side of Figure 2. This pattern supports the existence of a perceptual ceiling in tonal languages, a concept discussed by Xiao and Liu (2024), who showed that emotional prosody recognition in Mandarin is constrained by the need to preserve lexical tone contours.

Given that pitch in Mandarin must simultaneously convey lexical identity and emotional nuance, its functional load imposes stricter limits on prosodic manipulation. X. Liu et al. (2021) identified a perceptual threshold around 5–7 semitones for expressing surprise in Mandarin without distorting tone perception, which corresponds closely with our findings that moderate enhancements are most effective. Beyond this threshold, aggressive  $F_0$  scaling may compromise tone recognition, leading to reduced intelligibility or emotional plausibility. These results reinforce earlier conclusions by Wang et al. (2018) and P. Liu and Pell (2014), who reported that tonal language speakers compensate for pitch constraints by relying more on other cues such as duration and spectral features. Overall, Mandarin listeners appear more sensitive to pitch and energy modifications that violate tonal norms, suggesting a narrower viable range for expressive control in tonal TTS systems.

### 6.2.2 Linguistic Constraints and Pitch Functional Load

These findings can be explained in part by the functional load of pitch in each language. In English, pitch primarily fulfills intonational and pragmatic functions, such as indicating information structure, emotion, or discourse modality. As a result, pitch can be freely modulated for expressive purposes without risking semantic ambiguity.

Conversely, in Mandarin, pitch is lexically contrastive, distinguishing word meaning through tone categories (e.g., /mā/ “mother” vs. /mǎ/ “horse”). Thus, tone integrity is essential for intelligibility, and pitch deviations must remain within a narrow phonological range to avoid semantic confusion. While Mandarin speakers can and do use pitch excursions to express emotions such as surprise (X. Liu & Xu, 2016), they tend to do so in ways that preserve the underlying tone identity. Exceeding this safe prosodic range—as seen in the Strong condition—risks pushing the  $F_0$  contour beyond the tolerable phonetic envelope, leading to perceptual confusion or lexical ambiguity. This dual responsibility of pitch in Mandarin imposes a structural constraint on TTS systems, which must balance emotional salience with tonal accuracy.

These results support calls for language-specific tuning of prosody control systems (Guo et al., 2022), especially when extending expressive TTS into tonal language contexts. Our findings show that a one-size-fits-all strategy for pitch and energy enhancement is suboptimal and may even be counterproductive when deployed without tonal awareness.

### 6.2.3 Emotional Congruence and Keyword Alignment

The effectiveness of prosodic manipulation also depends on the semantic-pragmatic congruence of the manipulated keyword. English listeners may rely more on lexical stress patterns and pragmatic cues to interpret surprise. For example, the word “really?” in an upward rising contour naturally

signals incredulity in English prosody. Prosodic exaggeration here reinforces the affective reading without semantic interference.

In Mandarin, however, emotional interpretation is more context-sensitive and influenced by the interaction between lexical tone, sentence type, and discourse familiarity. Since Mandarin tones are lexically contrastive, the manipulation of  $F_0$  on emotionally salient keywords must preserve tonal identity to avoid ambiguity. If the manipulated keyword carries a rising or falling tone (e.g., Tone 2 or Tone 4), additional  $F_0$  modulation might distort the tone's category or even inadvertently shift its meaning (X. Liu et al., 2021). This could explain why some sentences in the Strong condition (e.g., Sentence 3) were rated lower than Moderate or Mild versions despite using identical enhancement protocols. Such cases reflect what Celle and Pélissier (2022) describe as “prosodic incongruence,” where acoustic emphasis fails to align with semantic or tonal expectations, leading to a breakdown in emotional communication.

Thus, emotional interpretation in Mandarin appears to require a more precise prosody-lexicon mapping. The success of keyword-level enhancement is not only dependent on the intensity of pitch and energy scaling, but also on its tonal compatibility and contextual fit. This highlights the need for tone-aware keyword selection mechanisms in future TTS systems aimed at tonal languages.

#### 6.2.4 Listener Expectations and Cultural Norms

Cross-linguistic variation in emotion perception may also reflect deeper cultural and cognitive expectations regarding vocal expressivity. English-speaking participants, accustomed to emotionally overt intonation in public communication (e.g., media, theater, and storytelling), may perceive dramatic prosodic cues as authentic and appropriate signals of affect (Ben-David et al., 2016; Dimos et al., 2015). This cultural familiarity with exaggerated intonation likely contributes to their positive reception of Strong prosody conditions in TTS output.

In contrast, Mandarin-speaking listeners may value emotional subtlety and prosodic naturalness, especially in formal or polite discourse contexts. Uthiraa and Patil (2023) and P. Liu and Pell (2014) found that tonal language speakers are more sensitive to deviations from expected prosodic norms, and may penalize speech that appears too exaggerated or unnatural—even if emotionally expressive. The notion of “appropriate emotional distance,” as explored in East Asian communication studies, further supports this observation: excessive vocal intensity may be interpreted not as emotional authenticity, but as stylistic incongruity or affective overreach.

In this regard, the Moderate enhancement level appears to offer a perceptual sweet spot for Mandarin listeners: sufficient pitch and energy elevation to suggest surprise, but restrained enough to preserve tonal identity and discourse naturalness. This balance between expressivity and intelligibility appears more delicate in tonal languages, reinforcing the importance of culturally and phonologically adaptive TTS control strategies (Guo et al., 2022; Luo et al., 2024).



### 6.3 Theoretical and Practical Implications

This study makes several theoretical and practical contributions to the design and evaluation of expressive TTS systems.

First, it demonstrates the interpretability of prosody control when grounded at the keyword level. Rather than relying on opaque style embeddings or global emotional tokens, the use of keyword-specific pitch and energy modulation ensures transparent and linguistically motivated control. This aligns with recent research trends emphasizing symbolic and controllable emotional synthesis mechanisms.

Second, the method is lightweight and modular. It does not require retraining or architectural modification of the TTS model. Instead, prosodic enhancement is achieved during inference by manipulating outputs of the variance adaptor, enabling flexible integration with existing FastSpeech2 pipelines. This makes it especially suitable for low-resource or modular deployment scenarios.

What’s more, the results indicate promising cross-linguistic generalizability. While keyword-level enhancement worked in both English and Mandarin, the observed differences underscore the importance of phonological awareness, especially in tonal languages. Therefore, the approach holds potential for developing phonology-sensitive but broadly applicable TTS control frameworks.

Finally, this approach has practical value for various real-world applications. Emotionally expressive, semantically grounded speech could enhance the quality of audiobook narration, dialogue agents, second-language learning systems, and assistive voice technologies by increasing listener engagement and emotional clarity.

### 6.4 Limitations

Despite these contributions, the study has several limitations that should be acknowledged:

First, the perception tests were based on a small set of manually selected sentences (five per language). Although semantically aligned and moderately long, this limited the syntactic and contextual variety of stimuli, potentially restricting generalizability.

Second, the number of participants was constrained. While perceptual trends were consistent, a larger and more diverse sample would be necessary to draw statistically robust conclusions and account for listener variability.

Third, the study focused solely on pitch and energy as prosodic features. Other cues such as duration, spectral tilt, and voice quality (e.g., breathiness or roughness) were not manipulated and could be explored in future research.

Finally, the Mandarin keyword selection process did not incorporate tonal categories or tone sandhi

rules, which may affect tone preservation and naturalness. This highlights the need for tone-aware detection in tonal language applications.

## 6.5 Future Directions

Building upon the findings of this study, several future research directions can be proposed to enhance the effectiveness and applicability of keyword-level prosody control in expressive TTS.

One important direction is the expansion of the emotional repertoire. While this study focused on surprise, future work could apply the same method to other emotional categories such as curiosity, sarcasm, or fear. Testing its effectiveness across a broader range of affective states would help assess the generalizability and versatility of the approach.

Another promising area involves the integration of tone-aware enhancement strategies, particularly for tonal languages like Mandarin. Because pitch and energy modifications may interfere with lexical tone realization, future systems should incorporate tonal constraints and phonological models to preserve intelligibility and tonal clarity.

A third avenue for exploration is the incorporation of additional prosodic dimensions. In addition to pitch and energy, future models could manipulate duration, spectral tilt, or voice quality to construct more complex and natural-sounding emotional speech. This would enable a more holistic form of expressive control.

Objective evaluation methodologies also warrant further development. The current study relied on subjective perceptual ratings; however, future experiments could include reaction-time measurements, physiological metrics (e.g., eye tracking, pupillometry), or neurocognitive tools (e.g., EEG, fMRI) to provide more rigorous validation of emotional perception.

Finally, future research should consider integrating keyword-level control with semantic understanding and emotion recognition modules. By enabling dynamic keyword selection based on context and speaker intent, such systems could support real-time, adaptive emotional synthesis. Moreover, interactive tools that allow users to manually select or adjust emotionally salient words could enhance customization in practical applications such as education, storytelling, or therapeutic communication.

## 7 Conclusion

This study investigated whether keyword-level pitch and energy enhancement can effectively increase the perception of surprise in synthetic speech, and whether this effect generalizes across typologically distinct languages—namely English and Mandarin Chinese. Grounded in recent advances in expressive text-to-speech (TTS) synthesis and motivated by the limitations of global emotional conditioning approaches, the study introduced a lightweight, interpretable mechanism for localized prosody control using FastSpeech2 and HiFi-GAN.

The results provide robust support for both research hypotheses. In response to RQ1, both forced-choice and scalar rating experiments demonstrated that prosodic enhancement—applied exclusively to semantically salient keywords—significantly increases listeners’ perception of surprise, without requiring retraining or model reconfiguration. This affirms the feasibility and perceptual salience of minimal, targeted interventions in the prosody space.

In addressing RQ2, the study uncovered meaningful cross-linguistic differences in emotional prosody perception. While English listeners exhibited a linear and consistent mapping between increased prosodic intensity and perceived surprise, Mandarin listeners showed a more nuanced, non-linear response pattern. This divergence reflects the distinct phonological roles of pitch and energy across languages: in English, pitch and energy modulation freely contributes to emotional expressivity, whereas in Mandarin, excessive pitch and energy scaling may interfere with lexical tone integrity, limiting the effectiveness of strong prosodic manipulation.

The findings contribute theoretically to the development of controllable, interpretable TTS systems by showing that keyword-level prosody control can bridge the gap between symbolic semantic understanding and acoustic expressivity. Practically, the approach offers a modular, low-resource-compatible method for enhancing affective clarity in speech synthesis, applicable to a range of domains including virtual assistants, audiobooks, language learning, and therapeutic applications.

Nonetheless, the study also acknowledges several limitations, including a restricted stimulus set, limited participant diversity, and a focus on only two prosodic parameters (pitch and energy). These constraints suggest avenues for future work, such as expanding the emotional repertoire, incorporating additional prosodic and voice quality features, applying tone-aware enhancement strategies in tonal languages, and exploring objective evaluation metrics beyond subjective perception.

In sum, this research demonstrates that prosodic modulation at the keyword level can enhance emotional perception in synthetic speech in a controllable and perceptually transparent way. By advancing a cross-linguistically aware framework for expressive speech synthesis, the study lays foundational work for next-generation TTS systems that are both emotionally intelligent and linguistically adaptive.

## References

- Asu, E. L., Sahkai, H., & Lippus, P. (2024, February). The prosody of surprise questions in estonian. *Journal of Linguistics*, 60, 7–27. doi: 10.1017/S0022226723000014
- Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. H. M. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1), 72–89. Retrieved from [https://doi.org/10.1044/2015\\_JSLHR-H-14-0323](https://doi.org/10.1044/2015_JSLHR-H-14-0323) doi: 10.1044/2015\_JSLHR-H-14-0323
- Celle, A., & Péliissier, M. (2022, January). Surprise questions in spoken french. *Linguistics Vanguard*, 8. doi: 10.1515/lingvan-2020-0109
- Diatlova, D., & Shutov, V. (2023). EmoSpeech: Guiding FastSpeech2 towards emotional text to speech. In *Proceedings of the 12th isca speech synthesis workshop (ssw2023)* (pp. 106–112). ISCA. Retrieved from <https://doi.org/10.21437/SSW.2023-17> doi: 10.21437/SSW.2023-17
- Dimos, K., Dick, L., & Dellwo, V. (2015). Perception of levels of emotion in speech prosody. In *Proceedings of the 18th international congress of phonetic sciences (icphs 2015)*. Glasgow, UK: The Scottish Consortium for ICPhS. Retrieved from <https://www.internationalphoneticassociation.org/icphsproceedings/ICPhS2015/Papers/ICPHS0756.pdf>
- Guo, Y., Du, C., & Yu, K. (2022, February). *Unsupervised word-level prosody tagging for controllable speech synthesis*. Retrieved from <https://arxiv.org/abs/2202.07200> (arXiv preprint) doi: 10.48550/arXiv.2202.07200
- Ikeda, M., & Markov, K. (2024, August). FastSpeech2 based japanese emotional speech synthesis. In *Proceedings of the 2024 IEEE International Symposium on Multimedia (ISM)* (pp. 1–5). IEEE. doi: 10.1109/IS61756.2024.10705252
- Kitahara, Y. (1988, November). Prosodic components of speech in the expression of emotions. *Journal of the Acoustical Society of America*, 84, 1338–1346. Retrieved from <https://doi.org/10.1121/1.2026592> doi: 10.1121/1.2026592
- Lee, Y., & Kim, T. (2019, May). Robust and fine-grained prosody control of end-to-end speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5911–5915). IEEE. Retrieved from <https://doi.org/10.1109/ICASSP.2019.8683501> doi: 10.1109/ICASSP.2019.8683501
- Lim, D., Jung, S., & Kim, E. (2022). *JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech*. Retrieved from <https://arxiv.org/abs/2203.16852> (Paper presented at Interspeech 2022) doi: 10.48550/arXiv.2203.16852
- Liu, P., & Pell, M. (2014, May). Processing emotional prosody in mandarin chinese: A cross-language comparison. In *Proceedings of speech prosody 2014* (pp. 95–99). ISCA. doi: 10.21437/SpeechProsody.2014-7
- Liu, X., & Xu, Y. (2016, May). Pitch perception of focus and surprise in mandarin chinese: Evidence for parallel encoding via additive division of pitch range.. doi: 10.21437/TAL.2016-28
- Liu, X., Xu, Y., Zhang, W., & Tian, X. (2021, July). Multiple prosodic meanings are conveyed through separate pitch ranges: Evidence from perception of focus and surprise in mandarin chinese. *Cognitive, Affective, & Behavioral Neuroscience*, 21, 1–12. doi: 10.3758/s13415-021-00930-9

- Lu, C., Lee, J., Wen, X., Lou, X., & Oh, J. (2023). The samsung speech synthesis system for blizzard challenge 2023. In *Proceedings of the 18th blizzard challenge workshop* (pp. 52–57). Retrieved from <https://doi.org/10.21437/Blizzard.2023-6> doi: 10.21437/Blizzard.2023-6
- Lu, C., Wen, X., Liu, R., & Chen, X. (2021). Multi-speaker emotional speech synthesis with fine-grained prosody modeling. In *Proceedings of the 2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5729–5733). IEEE. doi: 10.1109/ICASSP39728.2021.9413398
- Luo, X., Takamichi, S., Saito, Y., Koriyama, T., & Saruwatari, H. (2024, January). Emotion-controllable speech synthesis using emotion soft label, utterance-level prosodic factors, and word-level prominence. *APSIPA Transactions on Signal and Information Processing*, 13. Retrieved from <https://doi.org/10.1561/116.00000242> (In press) doi: 10.1561/116.00000242
- Makarova, V. (2000, October). Acoustic characteristics of surprise in russian questions. In *Proceedings of the 6th international conference on spoken language processing (icslp 2000)* (pp. 658–661). ISCA. doi: 10.21437/ICSLP.2000-621
- Mozziconacci, S. (2002). Prosody and emotions. In *Proceedings of speech prosody 2002* (pp. 1–9). ISCA. doi: 10.21437/SpeechProsody.2002-1
- Peters, S. A., & Almor, A. (2015, March). *Creating the Sound of Sarcasm*. Paper presented at the 61st Annual Meeting of the Southeastern Psychological Association (SEPA). Hilton Head, SC. Retrieved from <https://www.researchgate.net/publication/271839917> (Accessed via ResearchGate)
- Rao, K., Koolagudi, S. G., & Vempada, R. (2012, June). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, 16. doi: 10.1007/s10772-012-9172-2
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. In *Proceedings of the 9th international conference on learning representations (iclr)*. Retrieved from <https://openreview.net/forum?id=piLPYqxtWuA>
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In *Advances in neural information processing systems* (Vol. 32, pp. 3171–3180). Retrieved from <https://proceedings.neurips.cc/paper/2019/file/f63f65b503e22cb970527f23c9ad7db1-Paper.pdf>
- Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2018, 04). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In (p. 4779-4783). doi: 10.1109/ICASSP.2018.8461368
- Sorin, A., Shechtman, S., & Pollet, V. (2015, April). Coherent modification of pitch and energy for expressive prosody implantation.. doi: 10.1109/ICASSP.2015.7178905
- Uthiraa, S., & Patil, H. A. (2023). Analysis of mandarin vs english language for emotional voice conversion. In *Proceedings of the 25th international conference on speech and computer (specom 2023), part ii* (pp. 295–306). Berlin, Heidelberg: Springer-Verlag. Retrieved from [https://doi.org/10.1007/978-3-031-48312-7\\_24](https://doi.org/10.1007/978-3-031-48312-7_24) doi: 10.1007/978-3-031-48312-7\_24
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *9th isca workshop on speech synthesis workshop (ssw 9)* (p. 125).

- van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., ... Hassabis, D. (2018). Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning (icml 2018)* (Vol. 80, pp. 3918–3926). PMLR. Retrieved from <https://proceedings.mlr.press/v80/oord18a.html>
- Vlčková-Mejvaldová, J., & Horák, P. (2011). Prosodic parameters of emotional synthetic speech in czech: Perception validation. In C. M. Travieso-González & J. B. Alonso-Hernández (Eds.), *Advances in nonlinear speech processing* (Vol. 7015). Berlin, Heidelberg: Springer. Retrieved from [https://doi.org/10.1007/978-3-642-25020-0\\_22](https://doi.org/10.1007/978-3-642-25020-0_22) doi: 10.1007/978-3-642-25020-0\_22
- Wang, T., Lee, Y.-C., & Ma, Q. (2018, December). Within and across-language comparison of vocal emotions in mandarin and english. *Applied Sciences*, 8(12), 2629. Retrieved from <https://doi.org/10.3390/app8122629> doi: 10.3390/app8122629
- Xiao, C., & Liu, J. (2024). The perception of emotional prosody in mandarin chinese words and sentences. *Second Language Research*. Retrieved from <https://doi.org/10.1177/02676583241286748> (Advance online publication) doi: 10.1177/02676583241286748
- Yang, J., Bae, J.-S., Bak, T., Kim, Y.-I., & Cho, H. Y. (2021). GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis. In *Proceedings of interspeech 2021* (pp. 2202–2206). Retrieved from <https://doi.org/10.21437/Interspeech.2021-971> doi: 10.21437/Interspeech.2021-971
- Zhou, Q., Xu, X., & Zhao, Y. (2024). Tibetan Speech Synthesis Based on Pre-Trained Mixture Alignment FastSpeech2. *Applied Sciences*, 14(15), 6834. Retrieved from <https://doi.org/10.3390/app14156834> doi: 10.3390/app14156834

## Appendices

### A Audio Demonstrations

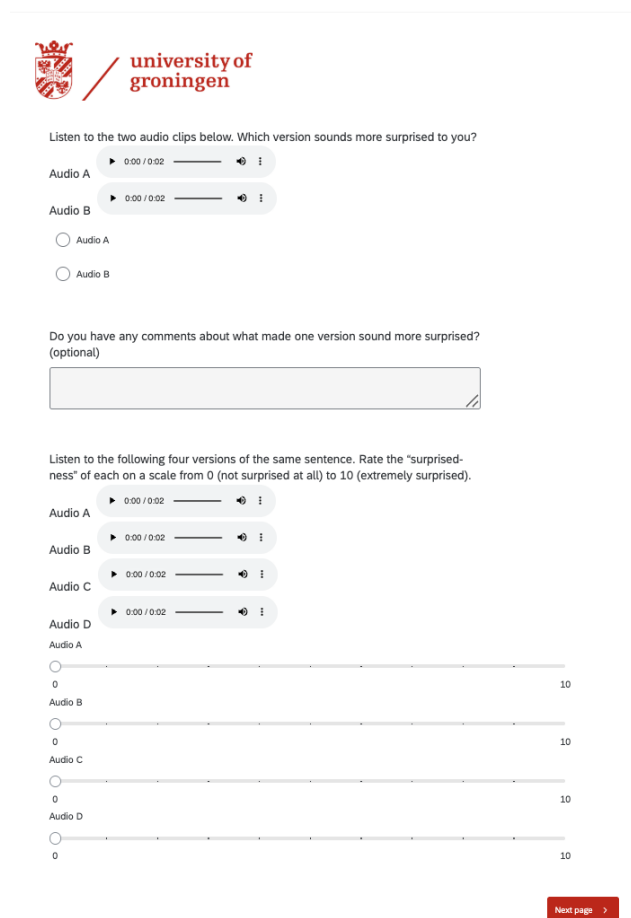
Pre-generated audio of the sample sentences used in the MOS test are available at <https://s-chen-rug.github.io/>


### B Source Code

For source code used in this study with adjustment based on original FastSpeech2 model and detailed documentation, visit the [https://github.com/S-CHEN-rug/Surprise\\_Perception](https://github.com/S-CHEN-rug/Surprise_Perception)


### C Listening Test Sample


Below is an example of the listening test used in this study. The survey consists of 10 test sentences for each language (English and Mandarin), each presented in the same format. Completing the survey takes approximately 20 minutes.



 university of groningen

Listen to the two audio clips below. Which version sounds more surprised to you?

Audio A 


Audio B 

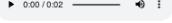
☐ Audio A

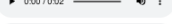
☐ Audio B

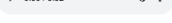
Do you have any comments about what made one version sound more surprised? (optional)

Listen to the following four versions of the same sentence. Rate the "surprisedness" of each on a scale from 0 (not surprised at all) to 10 (extremely surprised).

Audio A 

Audio B 

Audio C 

Audio D 

Audio A ☐ 0 10

Audio B ☐ 0 10

Audio C ☐ 0 10

Audio D ☐ 0 10

[Next page >](#)

Figure 3: Listening Test Survey Sample

## **D AI Usage Declaration**

I hereby affirm that this Master’s thesis was independently composed by myself. All work presented herein is my own, except where explicitly stated otherwise in the text. This thesis has not been submitted for any other academic degree or professional qualification, nor has it been published elsewhere. Where materials, ideas, or phrasing from other sources—whether printed, digital, or otherwise—have been referenced or adapted, these have been appropriately acknowledged.

During the preparation of this thesis, I made selective use of digital tools, including language models such as GPT-4o, to support clarity, coherence, and technical formulation in limited aspects of the writing process. Specifically, GPT-4o was used for sentence restructuring in Chapter 2, generating alternative explanations for technical concepts in Section 3.3, creating initial code function templates that are applied and discussed in Section 3.3, and summarizing background literature for preliminary review purposes. In addition, GPT-4o served as a source of inspiration during the early stages of code development, helping to explore possible structural patterns and implementation strategies. All AI-assisted content was thoroughly reviewed, validated, and significantly revised by me to ensure originality, accuracy, and academic integrity.

Shuyi CHEN / June 11, 2025