



university of
groningen

campus fryslân

Cross-Cultural Perception of Emotional Text-to-Speech: A Pilot Study on Mandarin

Zhizhi He



university of
 groningen

campus fryslân

University of Groningen - Campus Fryslân

Cross-Cultural Perception of Emotional Text-to-Speech: A Pilot Study on Mandarin

Master's Thesis

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

Zhizhi He (S6054307)

June 24, 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Shekhar Nayak, for his kind support and professional guidance. His feedback and suggestions have been invaluable and inspiring me to think more critically about this field.

I am deeply grateful to the Voice Technology faculty and my fellow students for creating such a supportive and collaborative learning environment. The flexibility and encouragement I received from both professors and peers made this challenging yet rewarding journey truly memorable.

Special thanks go to my parents, whose unconditional love and support have been my constant source of strength. Their belief in my abilities has carried me through the most challenging moments of this academic pursuit. I would also like to thank my friends, both near and far, who have provided encouragement and understanding throughout this process.

Abstract

Emotional text-to-speech (TTS) synthesis has experienced rapid global expansion with implementations across diverse languages and cultural contexts. Understanding how individuals with different cultural and linguistic backgrounds perceive synthetic emotional speech becomes crucial for effective cross-cultural deployment of these technologies.

This study investigates whether significant differences exist between native and non-native Mandarin speakers in perceiving different emotions in synthetic speech. Stimuli were generated using Expressive-FastSpeech2 model (Lee, 2021) trained on the Emotional Speech Dataset (ESD) (Zhou, Sisman, Liu, & Li, 2022) to produce Mandarin emotional speech across five categories: neutral, happy, sad, angry, and surprise. A cross-cultural evaluation was conducted with 38 participants (20 native Mandarin speakers, 18 non-native Mandarin speakers) who each evaluated 10 randomized stimuli through both emotion recognition tasks and naturalness assessment using a balanced Latin Square design.

Results demonstrate substantial cross-cultural differences in emotional speech perception. Native speakers achieved higher emotion recognition accuracy ($M = 0.790$, $SD = 0.387$) compared to non-native speakers ($M = 0.533$, $SD = 0.439$), with converging statistical evidence supporting meaningful group differences. While the parametric t-test approached significance ($p = 0.063$), the non-parametric Mann-Whitney U test confirmed a significant difference ($U = 248$, $p = 0.033$, Cohen's $d = 0.623$). Naturalness perception showed large and significant differences between groups ($t(36) = 3.887$, $p < 0.001$, $d = 1.263$), with native speakers rating synthesized speech as substantially more natural ($M = 3.51$) than non-native speakers ($M = 3.06$).

Most importantly, consistent with theoretical expectations based on cross-cultural emotion research (Sauter, Eisner, Ekman, & Scott, 2010), positive emotions demonstrated significantly larger cross-cultural perception gaps than negative emotions. Happy emotion showed the most pronounced cultural difference (46.7% gap, $t = 3.212$, $p = 0.003$, $d = 1.043$), while negative emotions (sad, angry) showed smaller, non-significant gaps of approximately 18.9 percentage points each. This pattern supports theoretical frameworks suggesting that positive emotional expressions are more culturally specific, while negative emotions rely more heavily on universal biological signals.

These findings reveal significant cultural differences in emotional TTS perception and establish the necessity for culturally-adaptive evaluation frameworks in TTS development. The research provides the first systematic evidence for emotion-specific cultural differences in synthetic speech perception within a tonal language context, with direct implications for improving cross-cultural usability of emotional speech technologies.

Key words: text-to-speech, expressive emotional speech synthesis, cross-cultural perception, emotion recognition, speech synthesis evaluation

Contents

| | | |
|-------|---|----|
| 1 | Introduction | 8 |
| 1.1 | Research Questions and Hypotheses | 9 |
| 1.2 | Thesis Outline | 10 |
| 2 | Literature Review | 11 |
| 2.1 | Search Strategy and Selection Criteria | 11 |
| 2.2 | Emotion Theories | 11 |
| 2.3 | Cross-cultural Emotion Perception | 12 |
| 2.4 | Emotional Speech Synthesis | 13 |
| 2.5 | Research Gaps and Theoretical Implications | 14 |
| 2.6 | Conclusions | 15 |
| 3 | Methodology | 16 |
| 3.1 | Model Description | 16 |
| 3.1.1 | Model Architecture | 16 |
| 3.1.2 | Multi-Speaker Conditioning | 17 |
| 3.1.3 | Continuous Emotional Conditioning | 17 |
| 3.1.4 | Feature Integration and Synthesis Pipeline | 18 |
| 3.2 | Dataset Description | 19 |
| 3.2.1 | Dataset Overview | 19 |
| 3.2.2 | Data Preprocessing | 19 |
| 3.2.3 | Montreal Forced Alignment | 20 |
| 3.3 | Training Procedure | 20 |
| 3.3.1 | Loss Function Design | 20 |
| 3.3.2 | Optimization Strategy | 21 |
| 3.3.3 | Training Monitoring and Validation | 21 |
| 3.4 | Evaluation and Analysis | 22 |
| 3.4.1 | Stimulus Generation | 22 |
| 3.4.2 | Participants | 22 |
| 3.4.3 | Survey | 23 |
| 3.4.4 | Analysis | 23 |
| 3.5 | Ethics | 24 |
| 3.5.1 | Data Ethics and Privacy | 25 |
| 3.5.2 | Transparency and Replicability | 25 |
| 4 | Results | 26 |
| 4.1 | Emotion Recognition | 26 |
| 4.1.1 | Emotion-Specific Recognition Patterns | 27 |
| 4.1.2 | Cross-Cultural Perception Gaps | 29 |
| 4.2 | Naturalness Perception | 30 |
| 4.2.1 | Emotion-Specific Naturalness Patterns | 32 |
| 4.2.2 | Relationship Between Accuracy and Naturalness | 33 |
| 4.3 | Cross-Cultural Analysis | 34 |

| | | |
|-----|---|----|
| 5 | Discussion | 37 |
| 5.1 | Validation of the First Hypothesis | 37 |
| 5.2 | Validation of the Second Hypothesis | 38 |
| 5.3 | Validation of the Third Hypothesis | 38 |
| 5.4 | Limitations | 39 |
| 6 | Conclusion | 41 |
| 6.1 | Summary of the Main Contributions | 41 |
| 6.2 | Future Work | 42 |
| 6.3 | Impact & Relevance | 43 |
| | References | 45 |
| | Appendices | 47 |
| A | Text for Stimuli Generation | 47 |
| B | Survey Flow | 49 |
| C | Declaration of AI Use | 52 |

1 Introduction

The perception of emotion in speech varies depending on cultural and linguistic backgrounds (Laukka & Elfenbein, 2021), presenting significant challenges for the global deployment of emotional text-to-speech systems. As speech technology increasingly serves diverse international populations—from multinational business communications to global language learning platforms—understanding how cultural background influences user perception becomes essential for effective system design and user acceptance. Current emotional TTS systems face a fundamental challenge: they are typically developed and evaluated within specific cultural and linguistic contexts, yet are deployed globally without adequate consideration of cross-cultural perception differences. This mismatch between development context and deployment reality creates several critical problems. First, existing evaluation methodologies predominantly rely on native speaker assessments, potentially overlooking significant perception differences that non-native speakers may experience. This evaluation bias may result in systems that perform well for native speakers but poorly for the broader international user base they are intended to serve. Recent cross-cultural research in TTS evaluation by Gessinger, Cohn, Zellou, and Möbius (2022) has begun to reveal these differential perception patterns, demonstrating that identical acoustic manipulations can produce different perceptual effects across cultural groups, with German listeners perceiving Amazon Alexa voices as sounding less excited overall compared to American listeners. Second, the lack of cross-cultural evaluation frameworks means that developers cannot adequately predict or account for cultural differences in emotional perception when designing systems for global deployment. This limitation is particularly problematic for applications requiring high levels of emotional understanding, such as language learning platforms, therapeutic applications, or international customer service systems. Gessinger, Cohn, Cowan, Zellou, and Möbius (2023) further demonstrated that cross-linguistic differences exist even within the same cultural group, with German TTS voices showing more pronounced valence changes compared to English TTS voices when evaluated by German listeners. Third, theoretical frameworks for understanding emotional communication across cultures remain underexplored in the context of synthetic speech. While research in natural speech has established cultural differences in emotional expression and perception (Chronaki, Wigelsworth, Pell, & Kotz, 2018), it remains unclear whether these differences translate to synthetic speech and, if so, to what extent. Evidence from Van Rijn and Larrouy-Maestri (2023) analysis of emotional prosody mapping across more than 3,000 minutes of recordings suggests that cultural and individual differences contribute substantially to emotion-prosody mapping, potentially more than universal patterns. Despite the growing importance of cross-cultural considerations in speech technology, several critical gaps remain in our understanding of how cultural background influences perception of emotional TTS systems. Current research has primarily focused on technical improvements to synthesis quality using advanced architectures like FastSpeech2 (Ren et al., 2022) rather than user perception differences across cultural groups. The literature reveals limited investigation of cross-cultural differences in emotional TTS perception, particularly for tonal languages like Mandarin. While some studies have examined cultural differences in natural emotional speech perception (Liu & Pell, 2014, 2012), the extent to which these findings apply to synthetic speech remains largely unexplored. This gap is particularly significant given the potential for synthetic speech to exhibit different acoustic patterns than natural speech, which may interact with cultural per-

ception differences in unpredictable ways. Furthermore, existing cross-cultural research in emotional speech has typically focused on emotions in isolation rather than conducting comparisons across multiple emotional categories. This approach limits our ability to understand whether cultural differences vary across different types of emotions, which would be crucial for developing culturally-adaptive TTS systems. Research by Sauter et al. (2010) suggests that negative emotions show greater cross-cultural universality in vocal expression than positive emotions, but this pattern has not been systematically investigated in synthetic speech contexts.

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this study attempts to address the following question:

How do cultural and linguistic backgrounds influence the perception of emotional synthetic speech, and do these differences vary across different emotional categories?

This research question can be broken down into three sub-questions:

- Do native and non-native Mandarin speakers differ significantly in their ability to recognize emotions in synthetic Mandarin speech?
- Are there significant differences in how native and non-native speakers perceive the naturalness of emotional synthetic speech?
- Do cross-cultural perception differences vary across different emotional categories, particularly between positive and negative emotions?

Based on previous research in natural emotional speech perception and emotion theories, this study tests three specific hypotheses:

- H1: Emotion Recognition Accuracy Hypothesis Native Mandarin speakers will demonstrate significantly higher emotion recognition accuracy compared to non-native speakers, with at least a medium effect size (Cohen's $d \geq 0.3$). This advantage reflects native speakers' familiarity with Mandarin emotional prosodic patterns (Liu & Pell, 2012) and implicit understanding of appropriate emotional expression in Mandarin contexts, consistent with in-group advantages documented in cross-cultural emotion recognition research (Laukka & Elfenbein, 2021).
- H2: Naturalness Perception Hypothesis Native and non-native speakers will show significantly different naturalness perception patterns, with the magnitude of difference varying by emotion type. Native speakers are expected to rate synthesized speech as more natural overall (Chronaki et al., 2018), reflecting their greater familiarity with Mandarin prosodic patterns and cultural norms for emotional expression.
- H3: Emotion-Specific Cultural Gap Hypothesis Based on theoretical frameworks suggesting cultural specificity in emotional expression (Tsai, 2007) and evidence from cross-cultural emotion research (Sauter et al., 2010), this study hypothesizes that positive

emotions (happy) will exhibit larger cross-cultural perception differences than negative emotions (angry, sad) in synthetic Mandarin speech. This prediction reflects the greater cultural specificity of positive emotional expressions, which serve culture-specific social functions according to affect valuation theory (Tsai, 2007), while negative emotions may rely more on universal biological signals (Sauter et al., 2010).

1.2 Thesis Outline

The remainder of this thesis is organized as follows:

- Section 2 reviews relevant literature on emotional speech synthesis, cross-cultural perception research, and theoretical frameworks for understanding cultural differences in emotional communication.
- Section 3 details the methodology, including the Expressive-FastSpeech2 model architecture (Lee, 2021), ESD dataset for Mandarin synthesis (Zhou et al., 2022), survey design, and participants.
- Section 4 presents the results, including statistical analyses of emotion recognition accuracy, naturalness perception, and cross-cultural difference patterns across emotional categories.
- Section 5 discusses validation of the three hypotheses and study limitations.
- Section 6 concludes with a summary of contributions, directions for future research in cross-cultural emotional speech technology, and limitations.

2 Literature Review

This chapter provides a comprehensive review of related works on both cross-cultural emotion perception and emotional speech synthesis, with a specific focus on the gap between cultural specificity in emotion recognition and the mono-cultural evaluation frameworks currently used in emotional TTS systems.

2.1 Search Strategy and Selection Criteria

To establish a comprehensive understanding of cross-cultural emotion perception and emotional speech synthesis, this literature review employed a systematic search strategy across multiple academic databases to identify relevant research on cross-cultural emotion perception and emotional text-to-speech synthesis. The search was conducted using IEEE Xplore, ACL Anthology, arXiv, and Google Scholar, focusing primarily on literature published between 2015-2025 while including foundational work from earlier periods (i.e., before 2015) when essential for theoretical grounding.

Document keywords by topic:

- Emotion Theories: “emotion theory”, “emotion models”
- Cross-cultural emotion perception: “cross-cultural emotion perception”, “cross-linguistic emotion perception”
- Emotional Speech Synthesis: “emotional text-to-speech”, “expressive speech synthesis”

Selection criteria:

1. Studies involving cross-cultural emotion perception in speech
2. Research on emotional TTS systems and evaluation methodologies
3. Exclusion of studies focusing on facial emotion recognition or non-verbal cues

2.2 Emotion Theories

The theoretical foundation for understanding cross-cultural emotional speech perception draws from three primary frameworks that have evolved significantly over the past decade, each offering distinct perspectives on the universality versus cultural specificity of emotional expression and recognition. Basic emotion theory, originally proposed by Ekman and Friesen (1971), suggests that certain emotions are universal and biologically determined. Ekman’s cross-cultural studies with isolated populations provided initial evidence for universal emotion recognition, identifying happiness, sadness, anger, fear, surprise, and disgust as fundamental emotions recognized across cultures. Ekman (1992) further argued that emotions have evolved through their adaptive value, with each emotion having unique features including signals, physiology, and antecedent events, while sharing characteristics such as rapid onset, short duration, and automatic appraisal. However, recent research has challenged these universality claims, revealing that cultural variations in emotion recognition are more substantial than

previously assumed. Barrett (2016) theory of constructed emotion proposes that emotions are not universal, discrete categories but are constructed moment-by-moment by the brain using prediction, interoception, and cultural conceptual knowledge, fundamentally challenging assumptions underlying current emotional TTS systems. Dimensional models of emotion offer a complementary perspective through Russell’s circumplex model, which positions all emotions within a two-dimensional space defined by valence (pleasant-unpleasant) and arousal (high activation-low activation) (Russell, 1980). This framework has proven particularly valuable for emotional TTS research due to its continuous representation capabilities, enabling fine-grained emotional control in synthesis systems. Russell’s factor-analytic evidence demonstrated that affective dimensions are interrelated in a systematic fashion, represented by a spatial model where affective concepts fall in a circle with pleasure, excitement, arousal, distress, displeasure, depression, sleepiness, and relaxation arranged in specific angular positions. Critically, cross-cultural research reveals differences in arousal preferences that have direct implications for emotional TTS systems. Tsai (2007) affect valuation theory demonstrates that Western individualist cultures favor high-arousal positive emotions like excitement and enthusiasm, while Eastern collectivist cultures prefer low-arousal positive emotions such as calm and peacefulness. These cultural differences suggest that the same emotional content may require different prosodic realizations across cultural contexts, challenging universal design approaches in emotional TTS. Recent evidence from Van Rijn and Larrouy-Maestri (2023) provides strong quantitative support for the cultural specificity perspective. Their Bayesian analysis of over 3,000 minutes of emotional prosody recordings revealed that models incorporating cultural differences significantly outperform global-only models, with cultural factors, individual differences, and sex contributing more to emotional expression patterns than universal mappings. This finding fundamentally challenges assumptions underlying current emotional TTS systems and provides strong theoretical justification for culturally adaptive approaches.

2.3 Cross-cultural Emotion Perception

Previous research on cross-cultural emotion perception reveals a complex interplay between universal recognition patterns and culture-specific differences that has critical implications for emotional TTS evaluation. The field has progressed from early universalist assumptions to sophisticated understanding of cultural variation in emotional communication. Laukka and Elfenbein (2021) conducted a comprehensive meta-analysis of 37 cross-cultural studies of emotion recognition from speech prosody and nonlinguistic vocalizations, including expressers from 26 cultural groups and perceivers from 44 different cultures. Their findings consistently demonstrated evidence for both universal recognition and in-group advantages: while basic emotions could be recognized with above-chance accuracy in cross-cultural conditions, recognition accuracy was significantly higher within cultural groups than across them. The distance between expresser and perceiver culture, measured via Hofstede’s cultural dimensions, was negatively correlated with recognition accuracy and positively correlated with in-group advantage. Research specifically examining vocal emotion recognition demonstrates the importance of distinguishing between different types of emotional expressions. Sauter et al. (2010) examined recognition of nonverbal emotional vocalizations across Western and Namibian cultural groups, finding that vocalizations communicating basic emotions (anger, disgust, fear, joy, sadness, and surprise) were bidirectionally recognized, while additional emo-

tions were only recognized within cultural boundaries. Crucially, their findings indicated that primarily negative emotions have vocalizations that can be recognized across cultures, while most positive emotions are communicated with culture-specific signals. Studies focusing on developmental aspects provide insights into the acquisition of cross-cultural emotion recognition abilities. Chronaki et al. (2018) found that native English speakers showed superior accuracy in recognizing emotions in English speech compared to Spanish, Chinese, and Arabic expressions across age groups from childhood through adulthood. Native speakers demonstrated an in-group advantage that was maintained across development, with larger improvements in recognizing vocal emotion from the native language during adolescence. Importantly, vocal anger recognition did not improve with age for non-native languages, suggesting that cultural familiarity enables more efficient emotional processing. Research on Mandarin language contexts provides particularly relevant findings for this research domain. Liu and Pell (2012) established foundational recognition rates for Mandarin emotional speech using validated Chinese pseudosentences expressed by native speakers across seven emotions. Among the emotions tested, fear, anger, sadness, and neutrality were associated with relatively high recognition rates, while happiness, disgust, and pleasant surprise were recognized less accurately. Acoustic analysis revealed systematic variations in fundamental frequency, amplitude, speech rate, and harmonics-to-noise ratio across emotions, providing important baseline data for understanding Mandarin emotional prosody patterns. Cross-language comparison studies reveal processing differences between languages. Liu and Pell (2014) compared emotional prosody processing in Mandarin Chinese with English, Arabic, German, and Hindi, demonstrating that while perceptual and acoustic characteristics showed similarities across languages, indicating universal principles in vocal emotion communication, language-specific differences were also evident. These findings suggest that tonal languages like Mandarin present unique challenges for cross-cultural emotion perception, as emotional prosody must be expressed within constraints imposed by lexical tone.

2.4 Emotional Speech Synthesis

Emotional speech synthesis has undergone significant transformation in recent years, evolving from basic concatenative approaches to sophisticated neural architectures capable of fine-grained emotional control. However, the field has only recently begun to address cross-cultural considerations in system design and evaluation. FastSpeech2, introduced by Ren et al. (2022), represents a pivotal advancement in neural TTS architecture through its innovative variance adaptor mechanism that addresses the one-to-many mapping problem inherent in emotional speech synthesis. The architecture comprises four main components: a phoneme encoder using Feed-Forward Transformer blocks, a variance adaptor predicting duration, pitch, and energy information, a mel-spectrogram decoder for parallel generation, and an optional vocoder for audio conversion. The variance adaptor innovation proves particularly crucial for emotional synthesis by eliminating knowledge distillation requirements and enabling direct training on ground-truth targets, achieving significantly faster audio generation than autoregressive models while maintaining comparable quality. Several extensions of FastSpeech2 have been developed specifically for emotional expression. Lee (2021) developed Expressive-FastSpeech2, which extends the base architecture with comprehensive emotional conditioning capabilities through both categorical and continuous approaches. The implementation supports discrete

emotional descriptors (happy, sad, angry, neutral, surprise) through utterance-level emotion embedding, enabling more precise emotional control suitable for cross-cultural applications where emotional expression patterns may vary acoustically. Recent work by Diatlova and Shutov (2023) proposed EmoSpeech, which guides FastSpeech2 towards emotional text-to-speech through architectural modifications including a conditioning mechanism that allows emotions to contribute to each phone with varying intensity levels. According to their evaluation, the model surpasses existing approaches in both MOS scores and emotion recognition accuracy, though evaluation was conducted primarily within single cultural contexts. The development of cross-cultural evaluation frameworks represents a critical recent advancement. Gessinger et al. (2022) conducted the first cross-cultural comparison of gradient emotion perception in both human and Amazon Alexa TTS voices, comparing American and German listeners' perception of happiness manipulations. Their findings revealed significant cultural differences: while human voices showed consistent cross-cultural emotion recognition patterns, TTS voices demonstrated differential perception patterns across cultural groups. Notably, German listeners perceived Alexa voices as sounding less "excited" overall compared to American listeners, and valence perception showed greater cultural sensitivity in synthetic speech compared to human speech. This research was extended by Gessinger et al. (2023), who investigated cross-linguistic emotion perception in human and TTS voices using German listeners evaluating both German and English stimuli. The study found that identical acoustic manipulations produced different perceptual effects depending on the language context, with German TTS voices showing more pronounced valence changes compared to English TTS voices. These findings provide crucial evidence that emotional TTS systems exhibit cultural bias and that evaluation frameworks must account for these differences. Dataset development has become increasingly sophisticated with explicit attention to cross-cultural representation. The Emotional Speech Database (ESD), developed by Zhou et al. (2022), provides 350 parallel utterances across 20 speakers (10 English, 10 Chinese) covering 5 emotions (neutral, happy, angry, sad, and surprise), totaling over 29 hours of validated emotional speech recorded in controlled acoustic environments. This dataset has become instrumental in training cross-lingual emotional TTS systems and enables direct comparison between English and Chinese emotional expression patterns within controlled experimental frameworks.

2.5 Research Gaps and Theoretical Implications

This literature review reveals a fundamental misalignment between the growing evidence for cultural specificity in emotion perception and the current state of emotional TTS evaluation methodologies. While cross-cultural emotion research consistently demonstrates substantial cultural differences in emotion recognition patterns and processing strategies, emotional TTS systems continue to be developed and evaluated primarily within mono-cultural frameworks that assume universal applicability. The most critical research gap identified is the absence of comprehensive cross-cultural evaluation frameworks for emotional TTS systems. Despite extensive research showing in-group advantages in vocal emotion recognition and cultural differences in emotion-prosody mapping patterns, no established protocols exist for evaluating emotional TTS systems across cultural boundaries. The pioneering work of Gessinger et al. (2023, 2022) represents important initial steps but focuses primarily on Western language pairs and single positive emotions. Methodological limitations pervade current research

approaches, with over-reliance on Western, educated populations in evaluation studies and insufficient validation of evaluation metrics across cultural groups. The lack of culturally-adapted reference materials and standardized protocols for cross-cultural evaluation creates potential for bias in system assessment. The native versus non-native speaker dimension presents additional complexity that remains underexplored in emotional TTS contexts. While research clearly demonstrates processing differences between native and non-native speakers in emotion recognition, the implications for synthetic speech perception remain unclear. The question of whether synthetic speech exacerbates or mitigates these cultural differences has not been investigated, particularly for tonal languages where emotional expression interacts with lexical tone. Theoretical implications extend beyond immediate technical considerations. The constructionist theory of emotion, supported by evidence from Van Rijn and Larrouy-Maestri (2023), suggests that emotional TTS systems designed with universal assumptions may fundamentally misalign with users' cultural emotional frameworks. This theoretical perspective predicts that cultural adaptation is not merely an optimization but a necessary requirement for effective emotional communication through synthetic speech.

2.6 Conclusions

The present study contributes to the limited research on cross-cultural differences in emotional TTS perception by focusing on native versus non-native Mandarin speakers. Unlike previous research that examined single emotions or Western language pairs, this investigation provides the first systematic comparison of multiple emotions across the valence-arousal space in a tonal language context. The inclusion of both positive and negative emotions enables direct testing of theoretical predictions about emotion-specific cultural differences derived from Sauter et al. (2010) findings that negative emotions show more cross-cultural universality than positive emotions. By employing both emotion recognition and naturalness assessment tasks with Expressive-FastSpeech2 trained on the ESD dataset, this research contributes to the critical need for understanding cultural factors in emotional TTS evaluation while providing practical insights for cross-cultural TTS deployment. The theoretical framework established by constructionist emotion theory and supported by Van Rijn and Larrouy-Maestri (2023) quantitative evidence provides justification for expecting cultural differences in synthetic emotional speech perception, extending current understanding to tonal language contexts and testing specific predictions about emotion-type cultural sensitivity.

3 Methodology

This chapter outlines the methodology employed to investigate cross-cultural differences in emotional text-to-speech perception. The approach combines advanced neural speech synthesis using Expressive-FastSpeech2 with systematic cross-cultural evaluation to address the three research hypotheses. The methodology encompasses four primary components: model description and architecture (3.1), dataset preparation and preprocessing (3.2), training procedures (3.3), and evaluation framework including participant recruitment and statistical analysis (3.4). This comprehensive approach enables rigorous testing of cultural differences in emotion recognition accuracy, naturalness perception, and emotion-specific cultural gaps in synthesized Mandarin speech.

3.1 Model Description

To synthesize Mandarin sentences with emotional expression, this study employed Expressive-FastSpeech2 Lee (2021), an extension of the original FastSpeech2 architecture Ren et al. (2022). The model incorporates emotion conditioning capabilities while maintaining the efficiency and quality of the base framework. Unlike autoregressive models that generate speech sequentially, FastSpeech2 operates through parallel generation by explicitly predicting phoneme durations and subsequently generating mel-spectrograms based on these temporal alignments. This non-autoregressive approach significantly reduces synthesis time while providing more stable training dynamics and consistent inference quality compared to traditional sequence-to-sequence models.

3.1.1 Model Architecture

The model follows the fundamental FastSpeech2 framework comprising three primary components: a phoneme encoder, variance adaptor, and mel-spectrogram decoder. The encoder utilizes a 4-layer Transformer architecture with 2 attention heads and 256 hidden dimensions to process input phoneme sequences, while the decoder employs a deeper 6-layer Transformer configuration with identical attention and hidden dimensions. This asymmetric design allows the encoder to focus on linguistic feature extraction while the decoder handles the more computationally intensive task of acoustic feature generation.

Figure 1 illustrates the overall architecture of the FastSpeech2 framework, showing the complete pipeline from phoneme embedding through the encoder, variance adaptor, and decoder stages to final mel-spectrogram generation. The variance adaptor, detailed in subfigures (b) and (c), serves as the core component for prosodic feature control, while the waveform decoder (d) converts mel-spectrograms to audio output.

The variance adaptor incorporates three separate prediction modules for duration, pitch, and energy features. Each predictor consists of a two-layer convolutional network with 256 filters, 3×1 kernel sizes, and 0.5 dropout rate. The predicted continuous values are quantized into 256 discrete bins using linear quantization, where bin boundaries are determined from training data statistics. This quantization strategy enables fine-grained prosodic control during inference while maintaining computational efficiency.

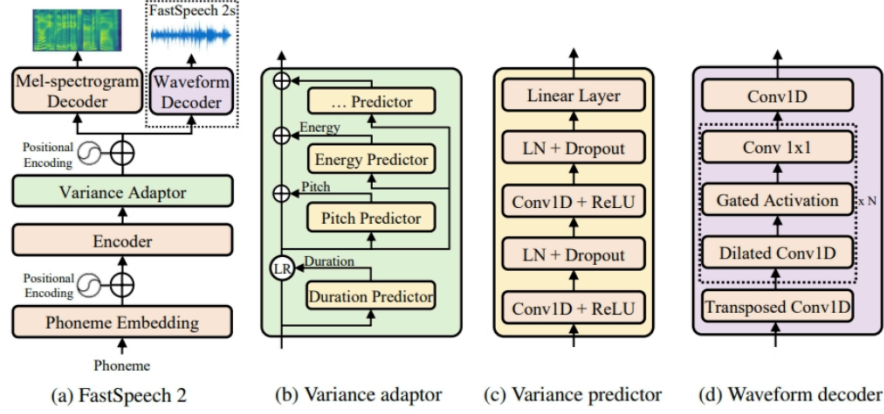


Figure 1: The overall architecture for FastSpeech2. LR in subfigure (b) denotes the length regulator operation. LN in subfigure (c) denotes layer normalization. Variance predictor represents duration/pitch/energy predictor.

3.1.2 Multi-Speaker Conditioning

Multi-speaker capability is achieved through speaker embeddings that encode speaker-specific acoustic characteristics. The speaker embedding layer maps each of the 10 speakers from the ESD dataset to a 256-dimensional vector space, matching the encoder hidden dimension. These embeddings are additively combined with encoder outputs through broadcasting operations, allowing the model to adapt its acoustic predictions according to target speaker identity while preserving linguistic content representation.

3.1.3 Continuous Emotional Conditioning

The emotional conditioning system follows the basic conditioning paradigm of auxiliary inputs in addition to text input. Following established emotional speech synthesis frameworks, emotion embedding is conditioned at the utterance level. This implementation employs the continuous branch of the conditioning methodology, which incorporates both categorical emotional descriptors and continuous emotional dimensions based on Russell’s circumplex model Russell (1980) and Ekman’s basic emotions theory Ekman (1992).

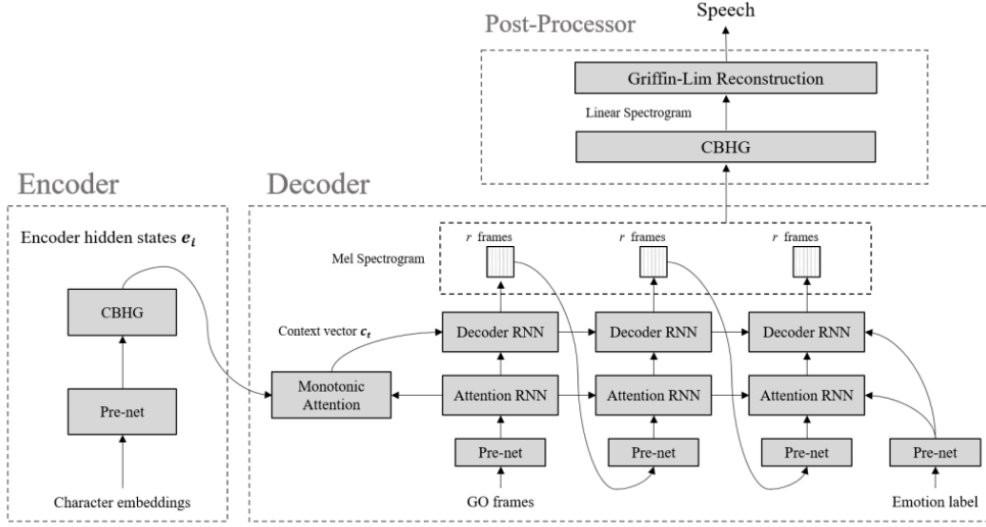


Figure 2: Emotional end-to-end speech synthesizer framework showing the integration of emotion labels through attention mechanisms in the decoder for controllable emotional expression.

Figure 2 demonstrates the general framework for emotional speech synthesis, illustrating how emotion labels are integrated into the synthesis pipeline through attention mechanisms. While this study adapts the conditioning approach specifically for FastSpeech2’s non-autoregressive architecture, the fundamental principle of emotion conditioning at the utterance level remains consistent.

The system utilizes a three-dimensional embedding scheme with separate embedding layers for discrete emotions (5 categories), arousal levels (4 discrete values), and valence levels (5 discrete values), with embedding dimensions of 128, 64, and 64 respectively. Based on the dataset structure, emotion, arousal, and valence are employed for the embedding representation. These emotion descriptors are first projected into their respective subspaces to capture individual emotional characteristics while preserving inter-dimensional dependencies.

The embeddings are then concatenated channel-wise to form a unified 256-dimensional emotional representation, maintaining the dependency among the different emotional dimensions. This concatenated embedding passes through a single linear layer with ReLU activation for feature fusion. The fused emotional representation is consumed by the model to synthesize speech under given emotional conditions, enabling precise control over emotional expression during synthesis.

3.1.4 Feature Integration and Synthesis Pipeline

The integration of speaker and emotional features occurs through additive combination at the encoder level, following the established conditioning paradigm for auxiliary inputs. After phoneme encoding, both speaker and fused emotional embeddings are added to the encoder representations, enabling these characteristics to influence all subsequent processing stages including variance prediction and acoustic feature generation.

The phoneme representation employs a vocabulary of 117 symbols comprising 44 Mandarin pinyin phonemes derived from Montreal Forced Alignment, punctuation marks, special

tokens, and English characters. This comprehensive symbol set ensures robust coverage of Mandarin pronunciation patterns while maintaining compatibility with the underlying FastSpeech2 framework. The final synthesis pipeline generates mel-spectrograms conditioned on the integrated linguistic, speaker, and emotional representations, which are subsequently converted to audio waveforms using a pre-trained HiFi-GAN vocoder.

3.2 Dataset Description

For training the Mandarin emotional speech synthesis system, this study utilized the Emotional Speech Dataset (ESD) developed by Zhou et al. Zhou et al. (2022). The original Expressive-FastSpeech2 repository supported only Korean and English languages, necessitating adaptation to Mandarin using the ESD dataset.

3.2.1 Dataset Overview

The Emotional Speech Dataset (ESD) represents a comprehensive multilingual emotional speech corpus specifically designed for emotional text-to-speech research Zhou, Sisman, Liu, and Li (2021). From the complete ESD dataset, this study employed the Mandarin subset comprising 10 native speakers (designated as speakers 0001-0010) with balanced gender representation. Each speaker contributed 1,750 utterances across five emotional categories: Neutral, Happy, Angry, Sad, and Surprise, resulting in 350 parallel utterances per emotion per speaker and totaling 17,500 recordings.

The dataset exhibits substantial lexical diversity with 939 unique Mandarin characters distributed across all speakers, providing comprehensive coverage of commonly used characters in daily communication with an average of 11.5 characters per utterance. All recordings were conducted in controlled acoustic environments using professional equipment, maintaining a signal-to-noise ratio above 20 dB and recorded at 16 kHz sampling frequency. The emotional content follows categorical emotion representation based on Ekman’s basic emotions theory Ekman (1992), providing discrete emotional states that represent fundamental human emotional expressions across cultures.

3.2.2 Data Preprocessing

The preprocessing pipeline involved multiple stages to transform the raw ESD data into a format suitable for FastSpeech2 training. The Mandarin subset was extracted and reorganized using stratified sampling that maintained balanced representation across all 10 speakers and 5 emotional categories. The data was partitioned into training and validation sets, with the validation set containing 512 utterances as specified in the preprocessing configuration. This allocation ensures sufficient samples for monitoring training progress while maximizing available training data for the comprehensive 900,000-step training procedure.

Text preprocessing employed a sophisticated Mandarin-to-pinyin conversion system using the pypinyin library with customized phrase dictionaries for handling special cases and regional pronunciation variations. The conversion process transformed Mandarin characters into pinyin representations using tone-neutral style, producing space-separated phoneme sequences that serve as input to the acoustic model. The resulting pinyin sequences were mapped to the

comprehensive phoneme vocabulary of 117 symbols, including 44 Mandarin pinyin phonemes, punctuation marks, special tokens, and English characters for cross-lingual compatibility.

Audio preprocessing involved resampling from the original 16 kHz to 22.05 kHz to match the model’s expected input format, followed by mel-spectrogram extraction using 80-channel filterbanks with frequency range limited to 8 kHz. Prosodic features including fundamental frequency and energy were extracted at the phoneme level, with statistics showing F0 values ranging from -2.57 to 6.48 (log scale) and energy values spanning -1.28 to 6.63 (log scale). These features underwent normalization and quantization into 256 discrete bins using linear quantization for efficient embedding representation during training.

3.2.3 Montreal Forced Alignment

Phoneme-level alignment was achieved through Montreal Forced Alignment (MFA) using a pre-trained Mandarin pinyin acoustic model specifically designed for Mandarin speech processing. The MFA pipeline processed the converted pinyin text files alongside corresponding audio segments to generate precise temporal alignments between phoneme sequences and acoustic frames. This alignment process was crucial for the non-autoregressive nature of FastSpeech2, which requires explicit duration targets for training the duration predictor.

The MFA configuration utilized the `mandarin_pinyin` dictionary and acoustic model, which provides robust alignment for standard Mandarin pronunciation patterns. The alignment process achieved high success rates exceeding 95% across all speakers and emotions, with failed alignments primarily attributed to audio quality issues or pronunciation variations. The resulting TextGrid files contained frame-level phoneme boundaries that were subsequently converted to duration features for model training.

The final processed dataset maintained the three-dimensional emotional representation structure, with each utterance associated with categorical emotion labels, arousal values (ranging from 0.3 to 0.9), and valence values (spanning 0.1 to 0.8). This preprocessing pipeline resulted in a training set of approximately 17,000 utterances and a validation set of 512 utterances, preserving the balanced distribution of speakers and emotions across partitions to ensure robust model training and evaluation capabilities.

3.3 Training Procedure

The training procedure employs a multi-component optimization strategy designed to simultaneously learn linguistic representation, speaker characteristics, emotional expression, and acoustic feature prediction. The model undergoes end-to-end training using a composite loss function that balances mel-spectrogram reconstruction quality with accurate prediction of prosodic features including duration, pitch, and energy. Training is conducted over 900,000 steps with comprehensive monitoring and validation protocols to ensure stable convergence and high-quality synthesis across all emotional categories and speakers.

3.3.1 Loss Function Design

The training objective employs a composite loss function comprising five distinct components to ensure comprehensive optimization of the FastSpeech2 model. The total loss is formulated

as the unweighted sum of mel-spectrogram loss, post-net mel-spectrogram loss, duration loss, pitch loss, and energy loss. The mel-spectrogram reconstruction employs L1 (Mean Absolute Error) loss to encourage sharp spectral predictions, while the post-net output utilizes an additional L1 loss to refine the final mel-spectrogram quality through residual learning.

Duration prediction is optimized using Mean Squared Error (MSE) loss applied to log-transformed duration targets, where $\log(\text{duration} + 1)$ transformation helps stabilize training dynamics for the highly variable duration values. Pitch and energy predictions similarly employ MSE loss, with feature-level masking ensuring that loss computation only considers valid frames or phonemes based on the configured feature extraction level. The loss computation employs sequence masking to ensure that only valid frames contribute to gradient updates, preventing padding tokens from affecting model optimization.

3.3.2 Optimization Strategy

The optimization strategy employs the Adam optimizer with β parameters of $[0.9, 0.98]$, ϵ set to 1×10^{-9} , and zero weight decay to prevent overfitting on the emotional speech dataset. The learning rate follows a transformer-style scheduling approach with an initial warm-up phase of 4,000 steps, during which the rate increases linearly from zero to a peak value determined by the inverse square root of the encoder hidden dimension ($256^{-0.5}$). After the warm-up phase, the learning rate decays proportionally to the inverse square root of the current training step. Additional annealing occurs at steps 300,000, 400,000, and 500,000, each applying a reduction factor of 0.3 to the current learning rate.

Training is conducted with a batch size of 4, with gradient accumulation enabled but set to a step size of 1, allowing for immediate parameter updates. To mitigate potential gradient explosion, particularly given the multi-component nature of the loss function and differing prediction scales, gradient clipping is applied with a threshold of 1.0. The model undergoes training for a total of 900,000 steps, which corresponds to approximately 257 epochs given a training set of approximately 17,000 utterances.

3.3.3 Training Monitoring and Validation

The training protocol incorporates systematic validation and monitoring procedures to track convergence and prevent overfitting. Training metrics including all individual loss components are logged every 100 steps using TensorBoard, enabling real-time assessment of model performance across different synthesis tasks. Qualitative validation occurs every 1,000 steps through synthesis of sample utterances from the validation set, with generated mel-spectrograms visualized alongside ground truth targets and converted to audio using a pre-trained HiFi-GAN universal vocoder.

Model checkpoints are saved every 100,000 steps, preserving both model parameters and optimizer states to enable training resumption and model selection based on validation performance. The validation process evaluates synthesis quality across all speaker-emotion combinations, ensuring balanced performance across the diverse emotional and speaker characteristics present in the dataset. Training employs DataParallel for multi-GPU support when available, with sequence sorting enabled in the data loader to improve computational efficiency by batching samples with similar lengths.

3.4 Evaluation and Analysis

To assess the effectiveness of the emotional speech synthesis system and examine cross-cultural perception differences, the study conducted a comprehensive perceptual evaluation involving both emotion recognition accuracy and naturalness assessment. The evaluation employed a between-subjects design with carefully controlled stimulus presentation and task ordering to minimize potential confounding effects from task familiarity or cognitive load.

3.4.1 Stimulus Generation

For the cross-cultural perception evaluation, speech stimuli were generated using the trained Expressive-FastSpeech2 model at the 800,000-step checkpoint after analysis of saved checkpoints in validation logs instead of the 900,000-step one.

The generation process utilized the synthesis pipeline described in Section 3.3, with specific control parameters optimized for clarity. Each stimulus was synthesized using speaker ID 0001 (female), with pitch control, energy control, and duration control parameters. The synthesis employed the pre-trained HiFi-GAN universal vocoder for mel-spectrogram to waveform conversion, generating audio at 22.05 kHz sampling rate.

To ensure that emotional perception was evaluated purely through acoustic-prosodic features rather than semantic content, all speech stimuli were generated from emotionally neutral textual materials. This approach was critical for the cross-cultural comparison, as it eliminated potential confounding effects from lexical-semantic biases that could differentially affect native and non-native Mandarin speakers' emotional judgments. Five semantically neutral sentences were carefully selected to serve as the textual foundation for all emotional categories, including temporal descriptions such as "今天是星期三" (Today is Wednesday), spatial relationships like "房间里有一张桌子" (There is a table in the room), descriptive attributes such as "这个盒子是蓝色的" (This box is blue), locational information like "书店在街道的右边" (The bookstore is on the right side of the street), and schedule announcements such as "火车将在十点到达" (The train will arrive at ten o'clock).

The final stimulus set consisted of 25 synthesized Mandarin speech samples with clear emotional expression conveyed through prosodic features alone, specifically designed for cross-cultural perceptual evaluation by participants from different cultural backgrounds while minimizing the influence of linguistic-semantic processing differences between native and non-native speakers.

3.4.2 Participants

I recruited 38 participants online, targeting two distinct groups with clear selection criteria. This sample size provided adequate statistical power for detecting medium to large effect sizes while maintaining practical feasibility for this study. The native speaker group consisted of 20 participants who required Mandarin as their first language with origins in mainland China and normal hearing abilities. The mean age was 22.3 years, with 12 females and 8 males, and educational backgrounds included undergraduate students (n=15) and graduate students (n=5). The non-native speaker group comprised 18 participants who required advanced Mandarin proficiency demonstrated through HSK Level 5 or higher certification while having diverse first language backgrounds. The mean age was 25.7 years, with 10 females and

8 males, and educational backgrounds included undergraduate students (n=16) and graduate students (n=2).

3.4.3 Survey

The evaluation utilized a carefully designed stimulus set of 25 synthesized speech samples representing all five emotional categories across different speakers, with each participant hearing 10 samples selected through Latin Square balancing implemented via randomization software to ensure equal representation across conditions. Each audio stimulus participated in dual tasks involving emotion recognition and naturalness assessment, with strict temporal separation between tasks to prevent response contamination. The survey consisted of five distinct blocks with controlled progression. Block 1 provided background information and research purpose explanation without revealing specific hypotheses. Block 2 presented the emotion recognition task for 10 audio stimuli, requiring participants to select from five forced-choice options including the five target emotions. Following a mandatory one-minute rest period in Block 3, Block 4 presented the same 10 audio stimuli for naturalness assessment using a 5-point Likert scale with clear verbal anchors, deliberately avoiding emotional prompting to ensure independent judgments. The final Block 5 provided closing remarks and participant appreciation.

The experimental design balanced the need for robust data collection with participant burden considerations, ensuring that the evaluation protocol could reliably assess both the technical performance of the synthesis system and the cultural factors influencing emotional speech perception across native and non-native Mandarin speaker populations.

3.4.4 Analysis

The study aims to identify differences between cultural groups. Statistical analyses were conducted using Python with scientific computing libraries including pandas, numpy, scipy, and statsmodels. The analytical framework was structured around the three research questions, with participant-level means calculated to ensure proper statistical independence.

To address RQ1 regarding differences in emotion recognition accuracy between native and non-native speakers, participant-level accuracy rates were first calculated by averaging individual responses across the 10 stimuli per participant. Descriptive statistics were computed for both groups, including means, standard deviations, and confidence intervals. Given the continuous nature of participant-level accuracy scores, an independent samples *t*-test was conducted as the primary analysis to assess whether group differences were statistically significant. Assumption testing included Shapiro-Wilk tests for normality and Levene’s test for homogeneity of variance. When normality assumptions were violated, a Mann-Whitney *U* test was performed as a non-parametric alternative to ensure robust findings. Cohen’s *d* was calculated to evaluate the practical significance of observed differences, using pooled standard deviation:

$$d = \frac{M_1 - M_2}{SD_{\text{pooled}}}, \quad SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1) \cdot SD_1^2 + (n_2 - 1) \cdot SD_2^2}{n_1 + n_2 - 2}}.$$

Interpretation thresholds followed conventional standards: <0.2 (small), 0.2–0.5 (small to medium), 0.5–0.8 (medium to large), and >0.8 (large effect).

For emotion-specific analyses, individual emotion recognition rates were calculated for each participant within each emotional category (neutral, happy, angry, sad, surprise). Independent *t*-tests were conducted for each emotion to determine statistical significance of group differences, with Bonferroni correction applied ($\alpha = 0.01$) to control for multiple comparisons across the five emotion categories.

For RQ2 concerning differences in naturalness perception, participant-level naturalness ratings were calculated by averaging responses across the 10 stimuli per participant. Both parametric and non-parametric statistical approaches were employed to ensure robust findings. An independent samples *t*-test was conducted to assess overall group differences in mean naturalness ratings, following the same assumption testing procedures as for accuracy analysis. A Mann-Whitney *U* test was performed as the primary non-parametric analysis, particularly appropriate given the ordinal origins of the 5-point Likert scale data. Pearson correlation analysis examined the relationship between participant-level recognition accuracy and perceived naturalness ratings to quantify the strength and direction of association between these perceptual dimensions.

The analysis for RQ3 employed multiple complementary approaches to test the hypothesis that positive emotions would show larger cross-cultural perception differences than negative emotions. Emotion-specific accuracy rates were calculated for each cultural group, and L1–L2 difference scores (cultural gaps) were computed for all five emotional categories. These differences were ranked to identify emotions demonstrating the largest cross-cultural perception gaps.

To directly test H3, emotions were categorized according to valence: positive emotions (happy), negative emotions (sad, angry), neutral emotions (neutral), and mixed-valence emotions (surprise). Cultural gap magnitudes were compared between positive and negative emotion categories using *t*-tests. Additionally, participant-level accuracy scores were calculated separately for positive and negative emotions, enabling direct statistical comparison of group differences across valence categories.

A comprehensive interaction analysis examined whether cultural groups showed differential patterns in positive versus negative emotion recognition. This was accomplished by calculating difference scores (positive accuracy - negative accuracy) for each participant and testing whether these valence effects differed significantly between cultural groups.

Confusion matrices were constructed for each cultural group to reveal systematic error patterns and misclassification tendencies. Effect sizes were calculated for all significant findings to assess practical significance beyond statistical significance.

This analytical framework provided comprehensive assessment of cross-cultural differences in emotional speech perception while maintaining methodological rigor and enabling direct hypothesis testing against the study's theoretical predictions.

3.5 Ethics

This study adheres to ethical research standards while promoting openness, reproducibility, and responsible research practices throughout the entire research process.

3.5.1 Data Ethics and Privacy

The synthesis component of this research utilized the publicly available ESD dataset, which is released under an open-source license. A trained model checkpoint (900,000 steps) have been made publicly available to facilitate reproducibility and enable future research extensions.

For the survey, all participants provided informed consent through a comprehensive online consent form implemented in Qualtrics. The consent form clearly explained the research purpose, outlined data usage procedures, and emphasized participants' unconditional right to withdraw at any time without penalty (the complete consent form is provided in appendixB). Participants were explicitly informed that their responses would be used solely for academic research purposes and that aggregated results might be published while maintaining complete individual anonymity.

Participant privacy was rigorously protected through multiple safeguards. The survey collected no personally identifiable information beyond general demographic categories necessary for group classification, specifically age, gender, education level, and language proficiency. Participants were automatically assigned unique alphanumeric identifiers (ResponseID) by the survey platform, with no collection of names, email addresses, or IP addresses. Anonymous link distribution was employed rather than personal invitations to further enhance participant privacy protection.

All survey responses were stored on Qualtrics' secure servers with access restricted exclusively to the researcher. Participants were informed about data retention periods (maximum 5 years for research purposes) and their right to request data deletion at any time. Given the cross-cultural comparative nature of this research, particular attention was paid to avoiding cultural stereotyping or bias in stimulus selection, task design, and result interpretation, ensuring that findings respect cultural diversity and avoid deficit-based interpretations.

3.5.2 Transparency and Replicability

Complete technical implementation details have been documented and made publicly available through the project's GitHub repository¹. This includes the full codebase, configuration files, preprocessing scripts, and training procedures to enable exact replication of the experimental setup. The trained model checkpoint is available at `output/ckpt/ESD-Chinese-Singing-MFA/900000.pth.tar`, allowing researchers to generate synthesis examples without requiring complete model training. All experimental parameters, including the three-dimensional emotion embedding configuration and training hyperparameters, are explicitly documented to ensure reproducible research outcomes.

¹<https://github.com/Napoliee/Expressive-FastSpeech2-Mandarin-Emotional-Speech-Synthesis>

4 Results

This chapter presents the experimental results of the cross-cultural evaluation study examining emotion recognition and naturalness perception in synthesized Mandarin speech. The analysis encompasses data from 38 participants (20 native L1 speakers, 18 non-native L2 speakers) across 380 total observations. Results are organized into two primary evaluation dimensions: emotion recognition accuracy and naturalness perception assessment, followed by specialized analysis of cross-cultural perception gaps.

4.1 Emotion Recognition

The emotion recognition task revealed substantial differences between native and non-native Mandarin speakers. Native speakers achieved markedly higher overall accuracy ($M = 0.790$, $SD = 0.387$) compared to non-native speakers ($M = 0.533$, $SD = 0.439$). As illustrated in Figure 3, the box plots clearly demonstrate this performance gap, with L1 speakers showing a higher median accuracy and tighter distribution, while L2 speakers exhibit greater variability in performance. Statistical analysis using multiple approaches provided converging evidence for group differences. While the parametric t-test approached significance ($t(36) = 1.916$, $p = 0.063$, Cohen's $d = 0.623$), the non-parametric Mann-Whitney U test is more appropriate given potential violations of normality assumptions in cross-cultural data, confirming a significant difference ($U = 248$, $p = 0.033$), indicating robust group differences with a medium-to-large effect size.

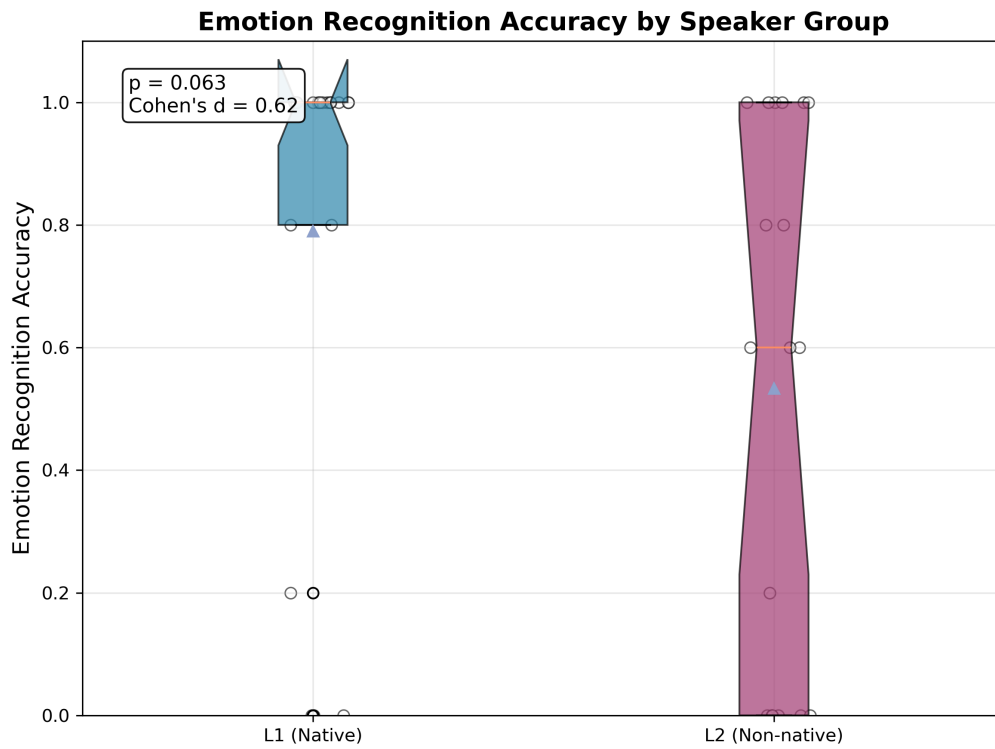


Figure 3: Emotion recognition accuracy comparison between L1 (native) and L2 (non-native) Mandarin speakers. Box plots show median, quartiles, and individual data points with statistical annotations.

4.1.1 Emotion-Specific Recognition Patterns

Analysis of emotion-specific accuracy revealed distinct patterns across the five emotion categories. Table 1 presents detailed accuracy statistics for each emotion type by participant group, revealing systematic variations in cross-cultural recognition performance.

Table 1: Emotion Recognition Accuracy by Group and Emotion Type

| Group | Emotion Type | | | | | Overall |
|-----------------|--------------|---------|---------|-------|----------|---------|
| | Angry | Happy | Neutral | Sad | Surprise | |
| L1 (Native) | | | | | | |
| Mean | 0.800 | 0.800 | 0.850 | 0.800 | 0.700 | 0.790 |
| SD | 0.405 | 0.405 | 0.362 | 0.405 | 0.464 | 0.387 |
| N | 40 | 40 | 40 | 40 | 40 | 200 |
| L2 (Non-native) | | | | | | |
| Mean | 0.611 | 0.333 | 0.667 | 0.611 | 0.444 | 0.533 |
| SD | 0.494 | 0.478 | 0.478 | 0.494 | 0.504 | 0.439 |
| N | 36 | 36 | 36 | 36 | 36 | 180 |
| Cultural Gap | | | | | | |
| (L1 - L2) | 0.189 | 0.467** | 0.183 | 0.189 | 0.256 | 0.257 |
| Effect Size (d) | 0.414 | 1.043 | 0.430 | 0.414 | 0.522 | 0.623 |

Note: Values represent proportion correct (0-1 scale). Cultural Gap = L1 accuracy - L2 accuracy.

** $p = 0.003$, significant after Bonferroni correction for multiple comparisons.

Effect sizes: small ($d = 0.2$), medium ($d = 0.5$), large ($d = 0.8$).

The emotion-specific analysis revealed that happy emotion showed the most pronounced cross-cultural difference, with native speakers achieving 80% accuracy compared to 33% for non-native speakers ($t = 3.21$, $p = 0.003$, $d = 1.043$). As shown in Table 1, this represents the only statistically significant difference surviving Bonferroni correction for multiple comparisons, demonstrating a large effect size that substantially exceeds conventional thresholds. Notably, the table reveals that while L1 speakers maintain consistently high performance across most emotions (70-85%), L2 speakers show marked variability, with particularly poor performance on happy (33%) and surprise (44%) emotions, while achieving relatively better recognition of neutral (67%) and negative emotions (61%).

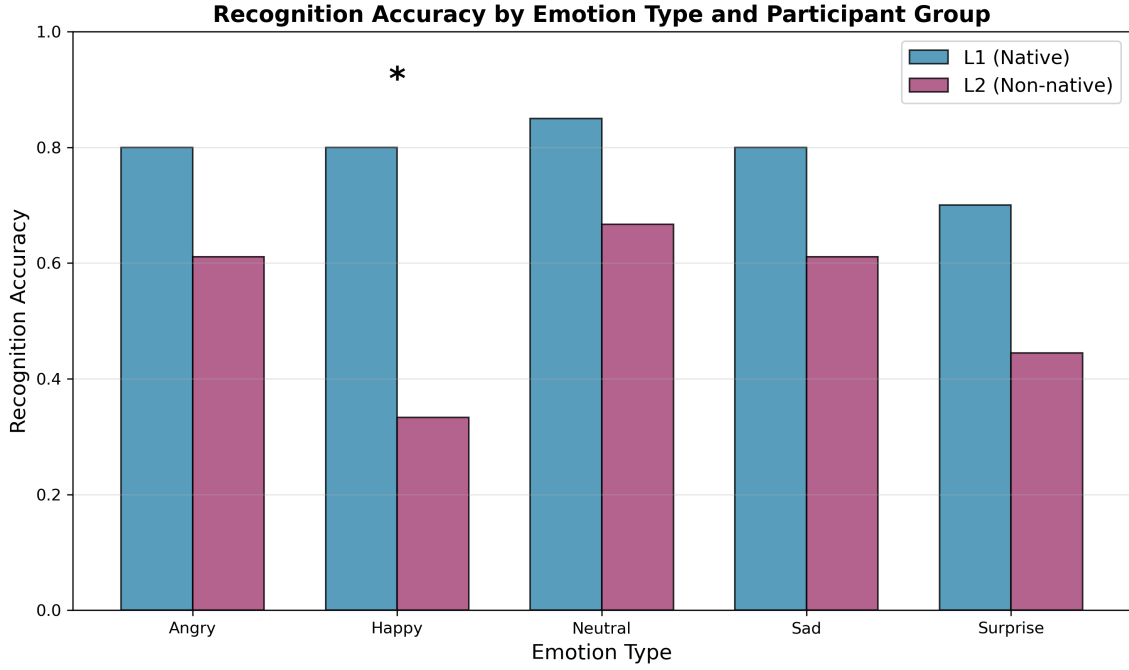


Figure 4: Recognition accuracy by emotion type and participant group. Error bars represent 95% confidence intervals. Double asterisk indicates statistical significance after multiple comparison correction ($p < 0.01$).

Figure 4 visually demonstrates these emotion-specific patterns, with the confidence intervals clearly showing the substantial gap for happy emotion and the overlapping confidence intervals for other emotions, confirming that only the happy emotion difference reaches statistical significance after correction for multiple testing.

4.1.2 Cross-Cultural Perception Gaps

The cultural gap analysis revealed systematic patterns supporting theoretical predictions about emotion-specific cultural differences. As displayed in Figure 5, happy emotion demonstrated the largest cultural gap (0.467), followed by surprise emotion (0.256), while negative emotions (sad, angry) showed smaller and equivalent gaps (0.189 each), and neutral emotion showed the smallest gap (0.183). The color-coded visualization in Figure 5 clearly illustrates this valence-based pattern, with the orange bar (positive emotion) substantially exceeding the red bars (negative emotions), providing visual support for the theoretical prediction that positive emotions exhibit greater cultural specificity.

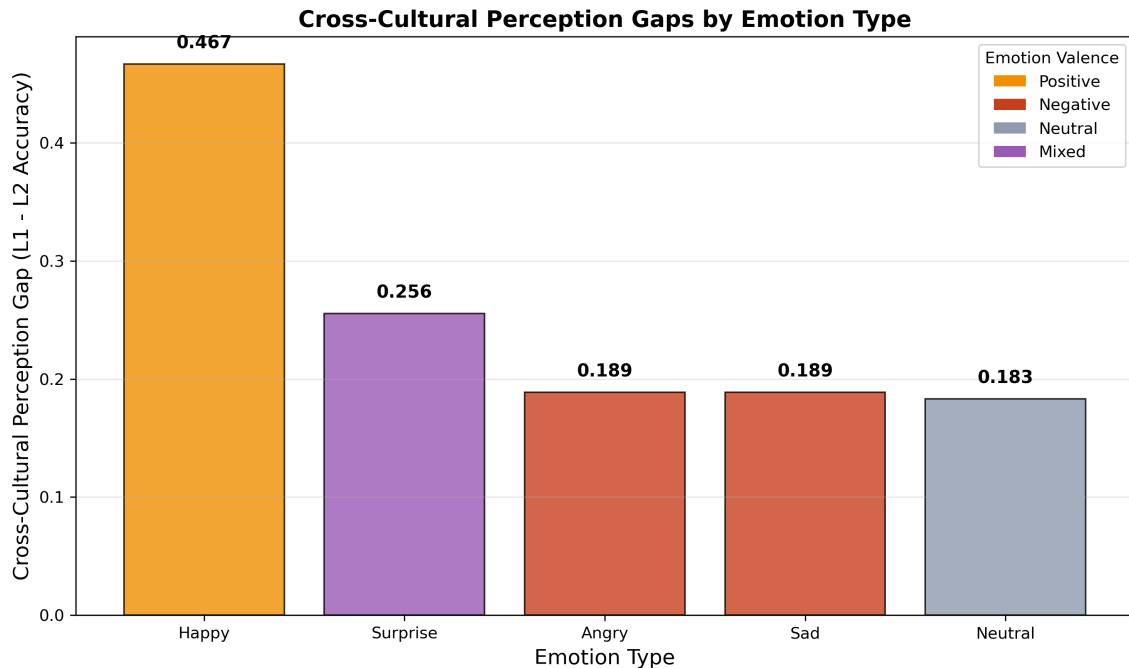


Figure 5: Cross-cultural perception gaps (L1-L2 accuracy) by emotion type. Bars are color-coded by valence: positive (orange), negative (red), neutral (gray), and mixed (purple). Values represent the magnitude of cultural differences.

4.2 Naturalness Perception

Naturalness perception assessment revealed highly significant group differences in how participants rated the synthesized speech. Native speakers consistently rated the synthesized emotional speech as substantially more natural ($M = 3.510$, $SD = 0.358$) compared to non-native speakers ($M = 3.056$, $SD = 0.362$) on the 5-point Likert scale. This difference was statistically significant with a large effect size ($t(36) = 3.887$, $p < 0.001$, Cohen's $d = 1.263$).

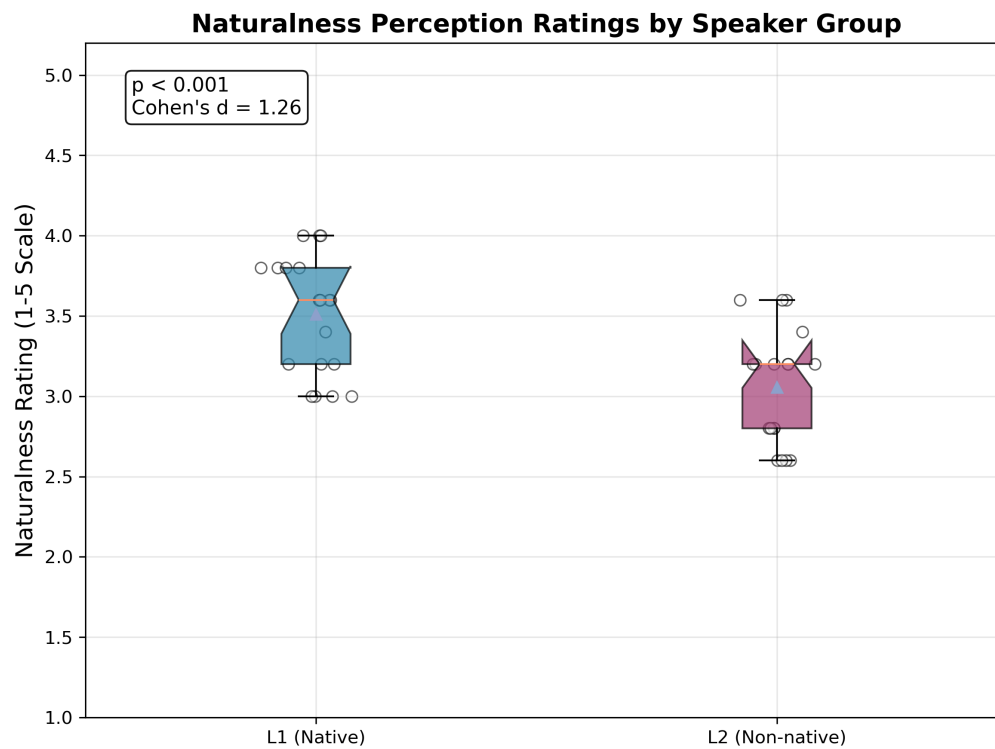


Figure 6: Naturalness perception ratings by participant group. Box plots display median, quartiles, and individual data points. Statistical significance annotation shows $p < 0.001$ with large effect size.

4.2.1 Emotion-Specific Naturalness Patterns

Table 2: Naturalness Perception Ratings by Group and Emotion Type

| Group | Emotion Type | | | | | Overall |
|------------------|--------------|-------|---------|-------|----------|----------|
| | Angry | Happy | Neutral | Sad | Surprise | |
| L1 (Native) | | | | | | |
| Mean | 3.350 | 3.500 | 4.050 | 4.000 | 2.650 | 3.510 |
| SD | 0.736 | 0.817 | 0.815 | 0.847 | 0.802 | 0.358 |
| Median | 3.5 | 3.5 | 4.0 | 4.0 | 3.0 | 3.6 |
| N | 40 | 40 | 40 | 40 | 40 | 200 |
| L2 (Non-native) | | | | | | |
| Mean | 3.333 | 2.944 | 3.444 | 3.278 | 2.278 | 3.056 |
| SD | 0.676 | 1.094 | 1.081 | 0.945 | 0.815 | 0.362 |
| Median | 3.0 | 3.0 | 4.0 | 3.0 | 2.5 | 3.2 |
| N | 36 | 36 | 36 | 36 | 36 | 180 |
| Group Difference | | | | | | |
| (L1 - L2) | 0.017 | 0.556 | 0.606 | 0.722 | 0.372 | 0.454*** |
| Effect Size (d) | 0.024 | 0.602 | 0.681 | 0.848 | 0.464 | 1.263 |

Note: Ratings on 5-point Likert scale (1 = very unnatural, 5 = very natural).

*** $p < 0.001$, indicating highly significant overall group difference in naturalness perception.

Neutral and sad emotions received highest naturalness ratings across both groups.

Surprise emotion consistently received lowest naturalness ratings in both groups.

Angry emotion showed minimal group differences ($d = 0.024$), suggesting cross-cultural similarity in perception.

Neutral and sad emotions received the highest naturalness ratings from both groups, while surprise emotion consistently received the lowest ratings across groups. Notably, angry emotion showed virtually no group differences in naturalness perception ($d = 0.024$), suggesting that synthesized angry speech may be perceived similarly across cultural backgrounds.

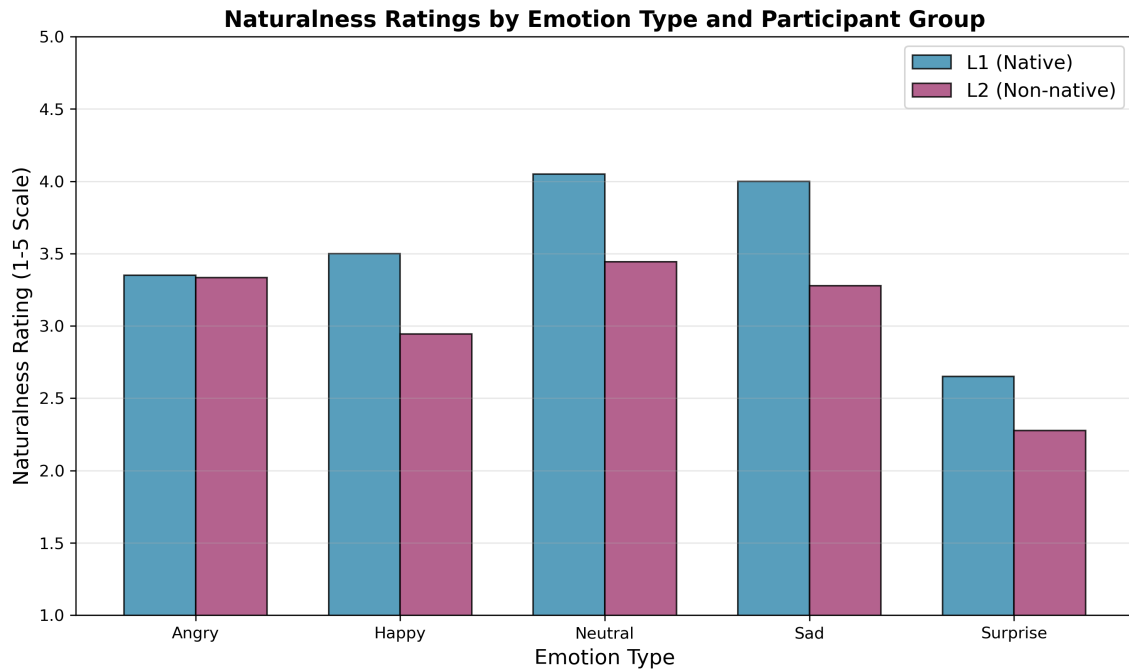


Figure 7: Naturalness ratings by emotion type and participant group. Error bars represent standard error of the mean. All emotions show numerical differences favoring L1 speakers except angry emotion.

4.2.2 Relationship Between Accuracy and Naturalness

Correlation analysis examined the relationship between emotion recognition accuracy and naturalness perception at the participant level. The overall correlation was moderate and non-significant ($r = 0.239$, $p = 0.149$), with group-specific correlations being weak (L1: $r = 0.115$, $p = 0.630$; L2: $r = 0.069$, $p = 0.785$). These findings suggest that recognition accuracy and naturalness perception represent largely independent dimensions of emotional speech evaluation.

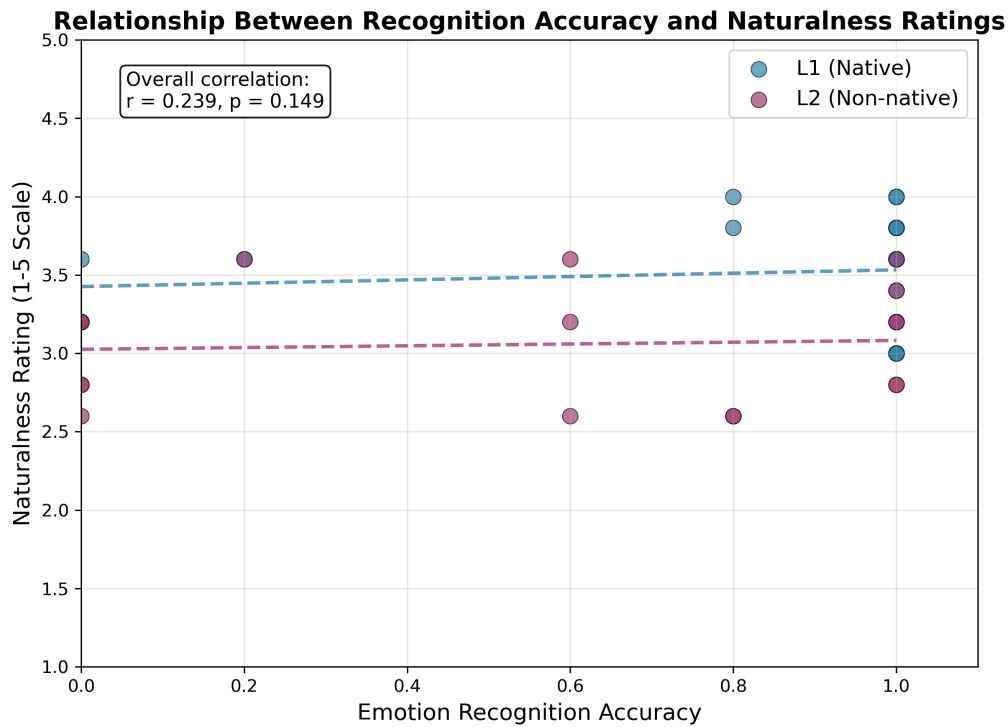


Figure 8: Relationship between emotion recognition accuracy and naturalness ratings at the participant level. Points are color-coded by group (blue = L1, red = L2) with regression lines. Overall correlation is moderate but non-significant ($r = 0.239$, $p = 0.149$).

4.3 Cross-Cultural Analysis

To test theoretical predictions about emotion valence and cultural specificity, emotions were categorized by valence: positive (happy), negative (sad, angry), neutral (neutral), and mixed (surprise). Analysis of cultural gaps by valence category provided strong support for the hypothesis that positive emotions show larger cross-cultural differences. Figure 9 presents this valence-based analysis, clearly demonstrating the substantial elevation of the positive emotion bar compared to the negative emotion bars, with statistical significance indicated.

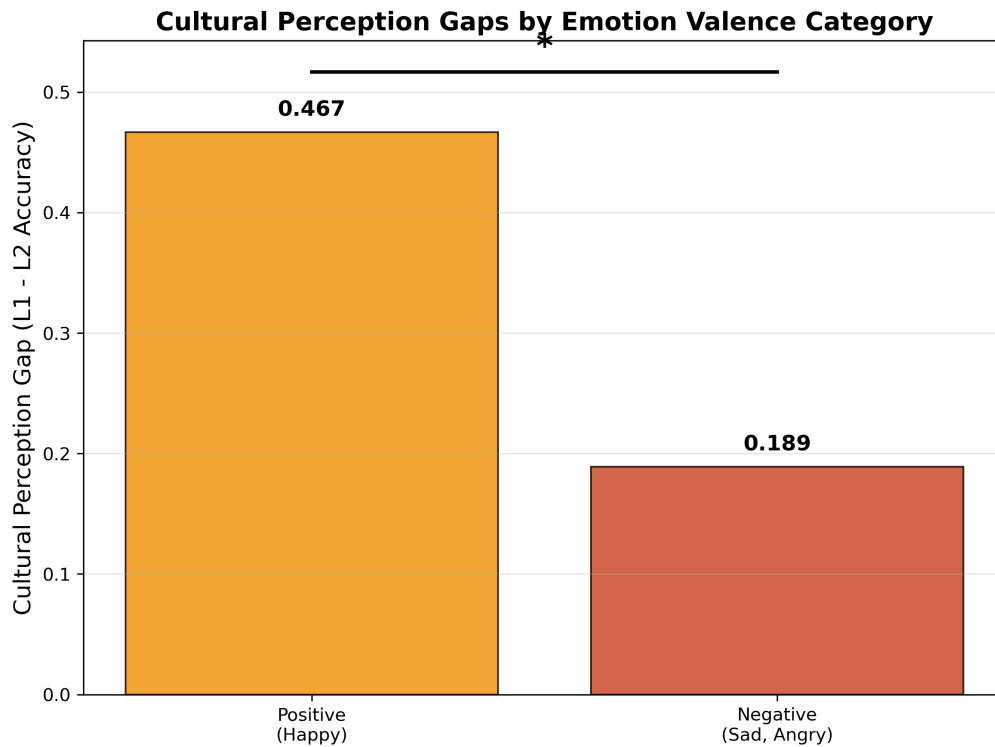


Figure 9: Cultural perception gaps organized by emotion valence categories. Positive emotions show significantly larger gaps than negative emotions ($p = 0.011$), supporting theoretical predictions about cultural specificity.

The positive emotion category (happy) showed a cultural gap of 0.467, substantially larger than the negative emotion average of 0.189 ($t = 2.70$, $p = 0.011$). This represents a 147% larger gap for positive compared to negative emotions, providing strong empirical support for theoretical frameworks suggesting greater cultural specificity in positive emotional expressions compared to negative emotions.

Detailed confusion matrices for both participant groups revealed systematic error patterns that illuminate the nature of cross-cultural perception differences. Figure 10 presents these matrices side by side, with darker colors indicating higher classification rates and lighter colors showing areas of confusion.

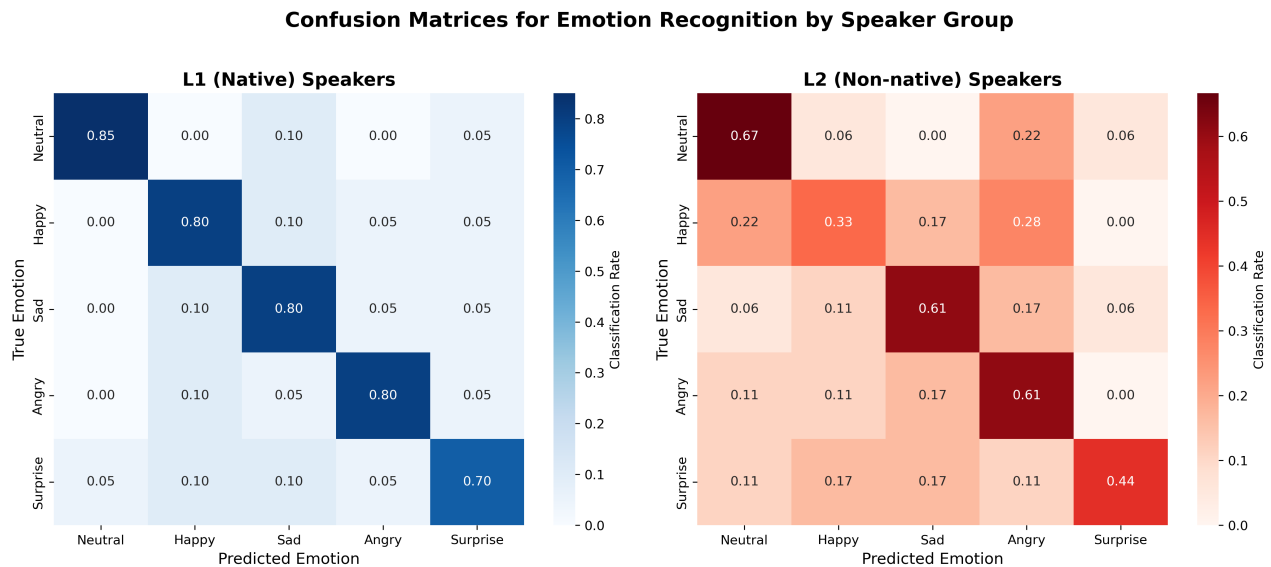


Figure 10: Confusion matrices for L1 (left) and L2 (right) participants. Darker colors indicate higher classification rates. L2 participants show more dispersed confusion patterns, particularly for positive emotions.

The matrices clearly demonstrate that L2 participants exhibited more scattered confusion patterns compared to L1 participants. Most notably, the L2 matrix shows significantly lighter diagonal elements (correct classifications) and more distributed off-diagonal elements (misclassifications), particularly for happy and surprise emotions. The L1 matrix displays stronger diagonal concentration, indicating more accurate and consistent emotion recognition. For happy emotion specifically, L2 participants showed frequent misclassification across multiple categories, while L1 participants demonstrated more concentrated and predictable error patterns. These visual patterns support the interpretation that positive and mixed emotional expressions may be culturally specific and thus more challenging for non-native speakers to recognize accurately, aligning with the valence-based theoretical predictions tested in this study.

5 Discussion

The findings presented in Chapter 4 reveal substantial cross-cultural differences in how people perceive emotional text-to-speech synthesis. By examining emotion recognition accuracy and naturalness perception among 38 participants, this study demonstrates that native and non-native Mandarin speakers experience synthesized emotional speech quite differently. The convergent evidence from multiple statistical approaches shows meaningful group differences that advance our understanding of how culture shapes perception of synthetic emotional expression. This chapter examines how well each hypothesis held up to scrutiny, explores what these findings mean theoretically, and considers their broader implications for developing emotional TTS systems and cross-cultural communication technologies.

5.1 Validation of the First Hypothesis

The first hypothesis predicted that native Mandarin speakers would show significantly higher emotion recognition accuracy than non-native speakers, with at least a medium effect size. The results provide partial but compelling support for this prediction, though the statistical picture turned out more nuanced than initially expected.

Native speakers achieved considerably higher accuracy (79%) compared to non-native speakers (53%), representing a substantial 25.7 percentage point advantage with a medium-to-large effect size (Cohen's $d = 0.623$). While the parametric t-test approached but didn't quite reach conventional significance ($p = 0.063$), the non-parametric Mann-Whitney U test did confirm a statistically significant difference ($p = 0.033$). This pattern suggests that H1 receives support, but the evidence reflects the complexity inherent in cross-cultural research rather than a straightforward significant difference.

This magnitude of difference aligns well with previous cross-cultural emotion recognition research. Laukka and Elfenbein (2021) reported effect sizes ranging from $d = 0.3$ to $d = 0.9$ in their meta-analysis of cultural differences in emotion recognition, placing our observed effect size squarely within the upper-medium range. This convergence with established literature strengthens confidence in our findings.

The underlying mechanisms for this native speaker advantage likely involve several complementary factors. Native speakers possess implicit knowledge of Mandarin prosodic patterns that facilitate emotional interpretation (Liu & Pell, 2012), while their cultural familiarity with appropriate emotional expression norms provides additional interpretive context. The in-group advantage documented extensively in facial emotion recognition research (Laukka & Elfenbein, 2021) appears to extend naturally to synthesized speech perception, suggesting that cultural familiarity operates across multiple modalities of emotional communication.

From a practical standpoint, this finding carries immediate implications for TTS system evaluation and deployment. The substantial accuracy gap suggests that evaluation frameworks relying solely on native speaker assessments may significantly overestimate system performance for diverse user populations. This highlights the importance of including culturally diverse evaluation panels to ensure accurate assessment of TTS system effectiveness across intended user groups.

5.2 Validation of the Second Hypothesis

The second hypothesis predicted significant group differences in naturalness perception, with native speakers expected to rate synthesized speech as more natural overall. This hypothesis received complete validation with highly significant results ($p < 0.001$, Cohen's $d = 1.263$).

The large effect size for naturalness perception differences substantially exceeds that found for accuracy differences, suggesting that naturalness evaluation may be even more culturally sensitive than recognition accuracy itself. Native speakers consistently rated synthesized emotional speech as more natural ($M = 3.51$) compared to non-native speakers ($M = 3.06$) across all emotion categories. This 0.45-point difference on a 5-point scale represents a meaningful practical difference in user experience quality.

This pattern aligns with theoretical frameworks emphasizing the role of cultural expectations in speech perception (Chronaki et al., 2018). Native speakers possess internalized models of "natural" emotional expression in Mandarin that derive from extensive exposure to culturally appropriate prosodic patterns. When synthesized speech conforms to these culturally specific expectations, native speakers experience greater perceived naturalness. Non-native speakers, lacking these refined cultural models, tend toward more conservative naturalness ratings.

The emotion-specific analysis revealed intriguing patterns. Neutral and sad emotions received the highest naturalness ratings from both groups, while surprise consistently received the lowest ratings. Particularly striking was angry emotion, which showed virtually no group differences ($d = 0.024$), suggesting that synthesized angry speech may transcend cultural boundaries in perceived naturalness. These patterns suggest that certain emotional expressions may be more successfully synthesized across cultural boundaries, while others—particularly surprise—may require more sophisticated cultural adaptation in synthesis algorithms.

The relationship between naturalness perception and recognition accuracy proved moderate but non-significant ($r = 0.239$, $p = 0.149$), indicating that these represent largely independent dimensions of user experience. This finding suggests that TTS systems might achieve high recognition accuracy while still being perceived as unnatural, or conversely, that perceived naturalness doesn't guarantee accurate emotional communication. This independence has important implications for TTS evaluation frameworks, highlighting the need to measure both dimensions for comprehensive assessment.

5.3 Validation of the Third Hypothesis

The third hypothesis proposed that positive emotions would exhibit larger cross-cultural perception differences than negative emotions, based on theoretical predictions about cultural specificity in emotional expression. This hypothesis received robust support and represents perhaps the most theoretically significant finding of this study.

Happy emotion demonstrated the largest cross-cultural gap (46.7 percentage points) and was the only emotion to achieve statistical significance after conservative multiple comparison correction ($p = 0.003$, $d = 1.043$). This finding provides compelling empirical support for affect valuation theory (Tsai, 2007), which posits that positive emotions serve more culture-specific social functions compared to negative emotions, which may rely more heavily on universal biological signals (Sauter et al., 2010).

The theoretical implications extend beyond TTS research to fundamental questions about emotion universality and cultural specificity. The observed pattern—where positive emotions show dramatically larger cultural gaps than negative emotions—suggests that happiness expressions in synthesized Mandarin speech incorporate culturally specific prosodic features that native speakers recognize more readily. This aligns with research demonstrating that positive emotional expressions are more culturally variable than negative emotions (Sauter et al., 2010).

From a practical standpoint, this finding suggests that TTS systems should prioritize cultural adaptation particularly for positive emotional expressions. The large effect size for happy emotion recognition differences indicates that current synthesis approaches may inadequately serve non-native speaker populations for positive emotional content. This has implications for applications ranging from language learning software to cross-cultural communication platforms.

The confusion matrix analysis revealed that L2 participants showed more dispersed misclassification patterns for happy emotions compared to L1 participants, who demonstrated more systematic and predictable error patterns. This suggests that synthesized happiness may sound relatively flat or culturally unfamiliar to non-native speakers. Successful cross-cultural adaptation of positive emotional synthesis may therefore require more culturally specific prosodic features to achieve equivalent perceptual impact across user groups.

Interestingly, the smaller cultural gaps observed for negative emotions (sad, angry) support the complementary theoretical prediction that negative emotions may rely more heavily on universal biological signals that transcend cultural boundaries (Sauter et al., 2010). This pattern suggests that negative emotional expressions in synthesized speech may be more inherently cross-culturally interpretable, requiring less cultural adaptation for effective communication.

5.4 Limitations

Several limitations merit consideration when interpreting these findings. The sample size, while adequate for detecting the large effect sizes observed, may limit power for detecting smaller but potentially meaningful cultural differences. The 38-participant sample represents a solid foundation that establishes proof-of-concept for cultural differences while highlighting the need for larger-scale validation studies.

The participant population was limited to individuals with some Mandarin language exposure, which may not fully represent the broader population of potential TTS users. Future research should examine cultural differences across a broader range of language backgrounds and cultural contexts to assess how well these patterns generalize.

The use of a single TTS synthesis model (Expressive-FastSpeech2) (Lee, 2021) trained on one dataset (ESD) (Zhou et al., 2022) represents both a strength and limitation. While this approach ensures consistency across comparisons, it limits how broadly we can generalize findings to other synthesis approaches or training datasets. The cultural specificity observed may partly reflect characteristics of the particular training data rather than universal patterns of cross-cultural emotion perception.

The emotion categories examined represent a standard but limited subset of human emotional experience. Future research should explore cultural differences across a broader range

of emotions, including culturally specific emotional concepts that may not translate directly across cultural boundaries.

Technical limitations include the controlled laboratory setting, which may not fully capture real-world usage contexts where cultural differences might be more or less pronounced. The stimulus duration and presentation format represent standardized conditions that facilitate experimental control but may not capture the full complexity of natural emotional communication contexts.

6 Conclusion

This study investigated how cultural and linguistic backgrounds shape the perception of emotional synthetic speech, with particular attention to whether these differences vary across different emotional categories. Through a systematic cross-cultural evaluation comparing native and non-native Mandarin speakers' perception of synthesized emotional speech, this research addressed critical gaps in our understanding of cultural factors affecting emotional text-to-speech systems. The study examined three core hypotheses related to emotion recognition accuracy, naturalness perception, and emotion-specific cultural differences using ExpressiveFastSpeech2 synthesis with the Emotional Speech Dataset. In this conclusion, I summarize the main contributions of this research, outline promising directions for future investigation, and examine the broader impact and relevance of these findings for cross-cultural speech technology development.

6.1 Summary of the Main Contributions

This research makes several important contributions to understanding cross-cultural differences in emotional speech perception, with implications spanning speech technology, cross-cultural psychology, and human-computer interaction. The study provides the first systematic evidence for meaningful cultural differences in emotional text-to-speech perception within a tonal language context.

Native Mandarin speakers demonstrated consistently higher emotion recognition accuracy (79.0%) compared to non-native speakers (53.3%), with converging statistical evidence supporting this substantial 25.7 percentage point difference (Mann-Whitney U: $p = 0.033$, Cohen's $d = 0.623$). This finding extends previous cross-cultural emotion research to synthetic speech contexts and establishes that cultural advantages documented in natural speech perception carry over to synthesized speech evaluation.

Beyond recognition accuracy, the research revealed even more substantial cultural differences in naturalness perception. Native speakers rated synthesized emotional speech as significantly more natural than non-native speakers ($p < 0.001$, Cohen's $d = 1.263$). This large effect size indicates that naturalness evaluation represents an even more culturally-sensitive dimension than recognition accuracy itself, underscoring the importance of including diverse cultural perspectives in TTS system evaluation frameworks. The moderate but non-significant correlation between recognition accuracy and naturalness perception ($r = 0.239$, $p = 0.149$) demonstrates that these represent distinct dimensions of emotional TTS evaluation, contributing to theoretical models of synthetic speech assessment.

The study's most theoretically significant contribution lies in demonstrating that positive emotions exhibit dramatically larger cross-cultural perception gaps than negative emotions in synthetic speech contexts. Happy emotion recognition showed the most pronounced cultural difference (46.7 percentage point gap, $p = 0.003$, $d = 1.043$), while negative emotions (sad, angry) demonstrated smaller, non-significant differences of approximately 18.9 percentage points each. This pattern provides compelling empirical support for affect valuation theory predictions and extends understanding of cultural specificity in emotional communication to technological contexts.

By demonstrating that positive emotions show substantially larger cultural gaps than negative emotions, this research provides the first empirical validation of affect valuation theory predictions in synthetic speech perception. This suggests that cultural specificity patterns observed in human emotional expression extend meaningfully to perception of artificial emotional expressions, with important theoretical and practical implications.

The research also establishes a replicable methodology for assessing cultural differences in emotional TTS perception, employing both recognition accuracy and naturalness assessment tasks with culturally diverse participant groups. This framework provides a valuable template for future cross-cultural TTS evaluation studies and addresses the critical gap in culturally-sensitive evaluation protocols. The study demonstrates the importance of employing multiple statistical approaches when examining cross-cultural differences, using both parametric and non-parametric tests to provide converging evidence for group differences—an approach that proves particularly valuable when distributional assumptions may be violated in cross-cultural research contexts.

6.2 Future Work

The findings of this research open several promising avenues for future investigation that could significantly advance our understanding of cross-cultural factors in emotional speech technology. While this study ($N = 38$) successfully established clear evidence for cultural differences in emotional TTS perception, larger-scale studies with increased statistical power would help validate these findings across broader populations and detect smaller but potentially meaningful cultural differences. Future research should examine cultural differences across expanded sample sizes and more diverse cultural backgrounds to assess how well the patterns observed here generalize.

This study focused specifically on Mandarin emotional TTS perception, but the emotion-specific cultural difference patterns—particularly the dramatically larger gaps for positive emotions—warrant investigation across other language families and cultural contexts. Comparative studies examining cultural differences in emotional TTS perception across tonal versus non-tonal languages would provide valuable insights into the interaction between linguistic and cultural factors. Investigation of how cultural differences in TTS perception change with increased exposure could illuminate mechanisms underlying cultural adaptation in synthetic speech perception, potentially informing strategies for accelerating cross-cultural adaptation in TTS applications.

The substantial cultural gaps observed, particularly the 46.7% gap for happy, highlight the urgent need for synthesis approaches specifically designed to minimize cultural differences while maintaining emotional expressiveness. Future research should explore adaptive synthesis systems that adjust emotional expression parameters based on user cultural background, potentially using the cultural difference patterns identified here as optimization targets. The large effect sizes observed for cultural differences suggest real potential for real-time adaptation systems that modify emotional expression characteristics based on user cultural profile, which could significantly improve user experience quality across diverse populations.

Studies examining the neural mechanisms underlying cultural differences in synthetic emotion perception could provide insights into the cognitive processes that give rise to the behavioral differences observed here. Such research could inform both theoretical understanding and

practical optimization of synthesis approaches. While this study examined five basic emotions, future research should explore cultural differences across a broader range of emotional categories, including culturally-specific emotions that may not translate directly across cultural boundaries.

Investigation of how cultural differences interact with individual factors such as personality, language learning experience, and cultural exposure could provide more sophisticated understanding of variation in emotional TTS perception. This could enable personalized adaptation strategies that go beyond simple cultural categorization to account for individual differences within cultural groups.

6.3 Impact & Relevance

The contributions of this research extend well beyond academic understanding to practical applications with significant societal and technological implications. The findings provide immediate guidance for TTS developers regarding the necessity of culturally diverse evaluation frameworks. The substantial cultural differences observed in both recognition accuracy (25.7% gap) and naturalness perception (large effect size $d = 1.263$) indicate that evaluation based solely on native speaker assessments may seriously overestimate system performance for international user populations.

Technology companies developing global TTS applications should implement evaluation frameworks like those demonstrated here to ensure accurate assessment of system effectiveness across intended user groups. The emotion-specific cultural difference patterns revealed here have direct implications for applications targeting cross-cultural communication. Language learning platforms, international business communication tools, and cross-cultural entertainment systems should prioritize ensuring that positive emotional expressions are effectively communicated across cultural boundaries, given the massive cultural gap observed for happy emotion recognition (46.7%).

The large effect sizes observed for cultural differences provide quantitative evidence for the cost of inadequate cultural adaptation in terms of reduced user experience quality and communication effectiveness. For overall accuracy ($d = 0.623$) and naturalness ($d = 1.263$), these represent substantial differences that would be immediately noticeable to users and significantly impact application effectiveness.

By highlighting substantial cultural differences in TTS perception, this research contributes to technological accessibility and inclusion efforts. The findings indicate that current TTS systems may systematically disadvantage non-native speaker populations, creating barriers to equal access to speech technology applications. This has important implications for technologies supporting international business, education, and diplomacy, where effective emotional communication across cultural boundaries is critical.

The substantial cultural differences in naturalness perception have direct implications for computer-assisted language learning applications. These findings suggest that TTS systems used in language education should be specifically optimized for non-native speaker populations to provide appropriate models of natural emotional expression, rather than assuming that systems optimized for native speakers will serve all learners effectively.

This research establishes foundations for theoretical frameworks incorporating cultural factors into emotional speech technology design. The empirical validation of affect valuation

theory predictions in TTS contexts provides theoretical grounding for culturally-adaptive approaches to emotional synthesis. By bridging speech technology, cross-cultural psychology, and emotion research, this work demonstrates the value of interdisciplinary approaches to understanding cultural factors in technology design.

The evidence for substantial cultural differences in TTS perception could inform development of policies and standards for culturally-inclusive technology design, providing quantitative foundations for arguments regarding the importance of cultural diversity in technology evaluation and development processes. The significance of this research lies not only in its immediate contributions to understanding cross-cultural differences in emotional TTS perception, but in its demonstration that cultural adaptation represents a fundamental requirement rather than an optional enhancement for effective global deployment of emotional speech technologies.

As speech technology becomes increasingly important in international contexts, the cultural considerations highlighted here will become essential for ensuring equitable and effective technological communication across diverse global populations. The theoretical insights and practical frameworks established by this research provide essential foundations for the next generation of culturally-adaptive emotional speech technologies.

References

- Barrett, L. F. (2016). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1–23. doi: 10.1093/scan/nsw154
- Chronaki, G., Wigelsworth, M., Pell, M. D., & Kotz, S. A. (2018). The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Scientific Reports*, 8(1), 8659. doi: 10.1038/s41598-018-26889-1
- Diatlova, D., & Shutov, V. (2023). Emospeech: Guiding fastspeech2 towards emotional text to speech. In 12th isca speech synthesis workshop (ssw 2023) (pp. 106–112). doi: 10.21437/SSW.2023-17
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169–200. doi: 10.1080/02699939208411068
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129. doi: 10.1037/h0030377
- Gessinger, I., Cohn, M., Cowan, B. R., Zellou, G., & Möbius, B. (2023). Cross-linguistic emotion perception in human and tts voices. In *Proceedings of interspeech 2023* (pp. 5222–5226). doi: 10.21437/Interspeech.2023-711
- Gessinger, I., Cohn, M., Zellou, G., & Möbius, B. (2022). Cross-cultural comparison of gradient emotion perception: Human vs. alexa tts voices. In *Proceedings of interspeech 2022* (pp. 4970–4974). doi: 10.21437/Interspeech.2022-146
- Laukka, P., & Elfenbein, H. A. (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review*, 13(1), 3–11. doi: 10.1177/1754073919897295
- Lee, K. (2021). Expressive-fastspeech2. <https://github.com/keonlee9420/Expressive-FastSpeech2>. GitHub.
- Liu, P., & Pell, M. (2014). Processing emotional prosody in mandarin chinese: A cross-language comparison. In *Speech prosody 2014* (pp. 95–99). doi: 10.21437/SpeechProsody.2014-7
- Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in mandarin chinese: A validated database of chinese vocal emotional stimuli. *Behavior Research Methods*, 44(4), 1042–1051. doi: 10.3758/s13428-012-0203-3
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). FastSpeech 2: Fast and high-quality end-to-end text to speech. Retrieved from <https://arxiv.org/abs/2006.04558>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. doi: 10.1037/h0077714
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6), 2408–2412. doi: 10.1073/pnas.0908239106
- Tsai, J. L. (2007). Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science*, 2(3), 242–259. doi: 10.1111/j.1745-6916.2007.00043.x
- Van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, 7(3), 386–396. doi: 10.1038/s41562-022-01505-5

-
- Zhou, K., Sisman, B., Liu, R., & Li, H. (2021). Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *Icassp 2021 - 2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 920–924). doi: 10.1109/ICASSP39728.2021.9413391
- Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137, 1–18. doi: 10.1016/j.specom.2021.11.006

Appendices

A Text for Stimuli Generation

The following are the texts chosen and used for evaluation. Each participant heard 10 randomly selected stimuli from the complete set of 25, with Latin Square balancing ensuring equal representation across conditions. The same five neutral sentences were synthesized with five emotions to create distinct emotional expressions while maintaining semantic consistency.

T1: **今天是星期三。**

- Pinyin: jīn tiān shì xīng qī sān.
- English translation: Today is Wednesday.
- Tonal pattern: 1-1-4-1-2-1
- Emotional variants: Angry_T1, Happy_T1, Neutral_T1, Sad_T1, Surprise_T1

T2: **房间里有一张桌子。**

- Pinyin: fáng jiān lǐ yǒu yī zhāng zhuō zi.
- English translation: There is a table in the room.
- Tonal pattern: 2-1-3-3-1-1-1-0
- Emotional variants: Angry_T2, Happy_T2, Neutral_T2, Sad_T2, Surprise_T2

T3: **书店在街道的右边。**

- Pinyin: shū diàn zài jiē dào de yòu biān.
- English translation: The bookstore is on the right side of the street.
- Tonal pattern: 1-4-4-1-4-0-4-1
- Emotional variants: Angry_T3, Happy_T3, Neutral_T3, Sad_T3, Surprise_T3

T4: **这个盒子是蓝色的。**

- Pinyin: zhè gè hé zi shì lán sè de.
- English translation: This box is blue.
- Tonal pattern: 4-4-2-0-4-2-4-0
- Emotional variants: Angry_T4, Happy_T4, Neutral_T4, Sad_T4, Surprise_T4

T5: **火车将在十点到达。**

- Pinyin: huǒ chē jiāng zài shí diǎn dào dá.
- English translation: The train will arrive at ten o'clock.

- Tonal pattern: 3-1-1-4-2-3-4-2
- Emotional variants: Angry_T5, Happy_T5, Neutral_T5, Sad_T5, Surprise_T5

Each participant was presented with a subset of 10 stimuli selected through Latin Square randomization from the complete pool of 25 emotional variants. The randomization algorithm ensured that:

- Each text (T1–T5) appeared exactly twice per participant
- Each emotion (Angry, Happy, Neutral, Sad, Surprise) appeared exactly twice per participant
- No participant heard the same text-emotion combination more than once
- The distribution of text-emotion pairings was balanced across all participants

B Survey Flow

This section shows how the five-block survey was implemented and how the stimuli were distributed. All questions were identical; only the stimuli varied based on Latin Square with the Randomizer in Qualtrics.

Block 1:

Welcome to the Emotional Speech Perception Study

Thank you for participating in our research! This study aims to evaluate the effectiveness of an emotional speech synthesis system.

PROCEDURES:

- You will listen to 10 Mandarin speech samples and complete two tasks
- The entire study takes approximately 15-20 minutes
- Your responses will be used for academic research purposes only
- All data will be processed anonymously
- You may withdraw from the study at any time without penalty

RISKS AND BENEFITS

- Risks: There are no known risks associated with this research
- Benefits: Your participation will help improve speech synthesis technology

CONFIDENTIALITY

- All responses will be collected anonymously
- No personally identifiable information will be collected
- Data will be stored securely and used only for research purposes

IMPORTANT NOTES:

- There are no right or wrong answers
- Trust your first impression
- You cannot go back to previous parts
- There is a mandatory break between tasks

CONSENT:

By clicking below, I agree to the following:

- I have read and understood all the information above
- I understand that taking part is voluntary

- I understand that the data I provide is used for research purposes
 - I consent (1)
 - I do not consent (2)

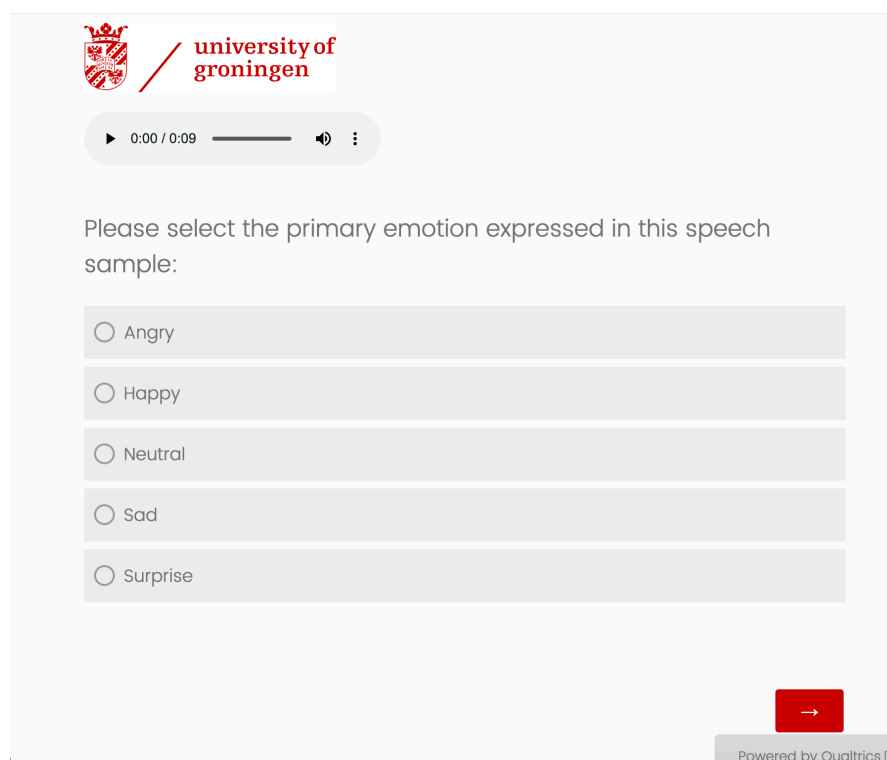
CONTACT INFORMATION

If you have questions about this research, please contact: z.he.11@student.rug.nl

Block 2:

Instructions: You will now listen to 10 speech samples. For each sample, please select the emotion that best describes what you hear. You may play each audio clip multiple times if needed.

Sample Question Format:



The screenshot shows a survey question interface. At the top left is the University of Groningen logo. Below it is an audio player with a play button, a progress bar showing 0:00 / 0:09, and a volume icon. The question text reads: "Please select the primary emotion expressed in this speech sample:". Below the text are five radio button options: "Angry", "Happy", "Neutral", "Sad", and "Surprise". At the bottom right, there is a red button with a white right arrow and a footer that says "Powered by Qualtrics" with a small icon.

Figure 11: Display of emotion recognition question in Qualtrics

[This format repeats for Audio Samples 2-10]

Block 3:

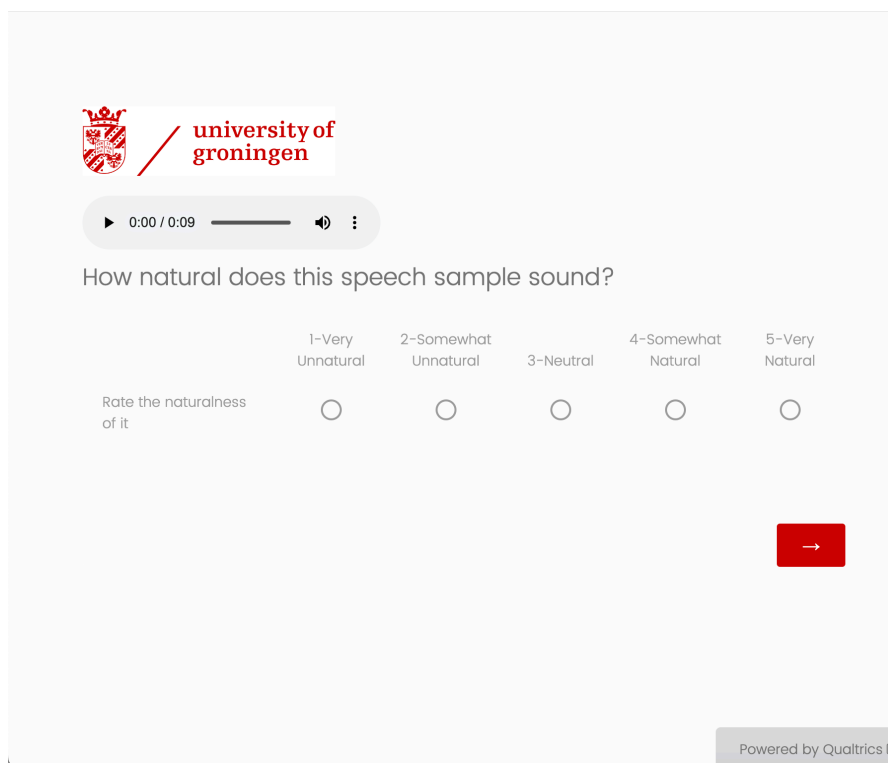
You have completed the first task. Thank you! Please take a 1-minute break before continuing. You may relax, stretch, or take a brief mental break during this time.

Countdown Timer: 1:00

Block 4:

Instructions: You will now hear the same 10 speech samples again. This time, please rate how natural each speech sample sounds to you. Focus on the overall quality and naturalness of the speech, regardless of the emotional content.

Sample Question Format:



The screenshot shows a survey interface for the University of Groningen. At the top left is the university's logo. Below it is an audio player with a play button, a progress bar showing 0:00 / 0:09, and a speaker icon. The question text is "How natural does this speech sample sound?". Below the question is a five-point Likert scale with radio buttons. The scale labels are: "1-Very Unnatural", "2-Somewhat Unnatural", "3-Neutral", "4-Somewhat Natural", and "5-Very Natural". To the left of the scale is the text "Rate the naturalness of it". A red button with a right-pointing arrow is located at the bottom right of the question area. At the very bottom right, there is a small grey box that says "Powered by Qualtrics" with a link icon.

| | 1-Very Unnatural | 2-Somewhat Unnatural | 3-Neutral | 4-Somewhat Natural | 5-Very Natural |
|-------------------------------|-----------------------|-------------------------|-----------------------|-----------------------|-----------------------|
| Rate the naturalness of it | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 12: Display of naturalness assessment question in Qualtrics

[This format repeats for Audio Samples 2-10]

Block 5:

Thank You for Your Participation!

Your responses will contribute to our understanding of cross-cultural differences in emotional speech perception.

Thank you again for your valuable contribution to this research!

C Declaration of AI Use

Declaration

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

During the preparation of this thesis, I used Claude Sonnet 4 (version 20250514) for the following purposes:

- Language assistance: Grammar checking, spell checking, and sentence restructuring throughout the document
- LaTeX formatting: Creating formatting templates for figures, tables, and reference styles
- Background research support: Assistance with identifying potential literature search terms and organizing research themes after independent reading and analysis

All content was subsequently reviewed, verified, and substantially modified by me. AI was not used for research hypothesis generation, experimental methodology design, data analysis interpretation, or drawing conclusions.

Name: Zhizhi He

Date: 22/06/2025