university of groningen / campus fryslân

Manipulating Acoustic Correlates for Vocal Persona Transition: From Neutral to Friendly

Chenyi Lin

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Manipulating Acoustic Correlates for Vocal Persona Transition: From Neutral to Friendly

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
PhD candidate Phat Do (Voice Technology, University of Groningen)
with the second reader being
(Voice Technology, University of Groningen)

Chenyi Lin (S5664713)

June 10, 2024

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Phat Do, for his invaluable technical assistance and for providing insightful suggestions and innovative ideas throughout the course of this thesis. His guidance and expertise have been instrumental in shaping the direction and quality of my work.

I am also profoundly grateful to Matt Coler for his early contributions to this thesis. His creative ideas and support during the initial stages were crucial in laying a strong foundation for this research. His encouragement and advice have greatly influenced the development of my thesis.

Finally, I would like to express my heartfelt thanks to my family and friends for their unwavering support and encouragement. Their belief in me and their continuous emotional support have been a source of strength and motivation, helping me to persevere through the challenges of this academic journey.

Abstract

The concept of vocal persona, reflecting the identity or character perceived through an individual's voice, exhibits dynamic variability as it adapts to various social contexts. Understanding the dynamic shifts of vocal persona not only enriches the expressivity and personalization of Text-to-Speech (TTS) systems but also holds potential for enhancing user engagement across applications. This adaptability is crucial for the effectiveness of TTS systems yet remains less explored, particularly in the realm of attitudinal nuances such as synthesizing speeech with a friendly attitude. The conventional method of synthesizing friendly speech involves training TTS models with datasets specifically containing friendly attitudes. However, given the limitations of available speech datasets, which predominantly lack diverse attitudinal tones, our study employed specific acoustic manipulations (namely alterations in pitch, duration, and energy) in neutral speech data to facilitate the perceptual transition of vocal personas from neutral to friendly in Mandarin Chinese TTS, using the FastSpeech2 framework. We examined the individual and combined effects of these acoustic features on enhancing the friendliness of synthesized speech. Through controlled experimental setups, our research quantified these perceptual shifts using identification accuracy and mean opinion scores (MOS).

Based on the findings of F. Chen, Li, Wang, Wang, and Fang (2004) and Li, Chen, Wang, and Wang (2004), we anticipated that increasing the mean pitch of a neutral voice alone will significantly influence friendliness perception. Moreover, integrating it with shorter phone duration and slightly raised energy was expected to further optimize the perception of friendliness. However, our study revealed that neither modulation of pitch alone nor alterations in pitch, duration, and energy together achieved a significant perceptual shift towards friendliness. This may suggest a limited effect of acoustic cues alone in friendliness perception and may require further investigation into the effectiveness of acoustic manipulations in synthesizing friendly speech. Despite the unsuccessful perceptual transition, this exploration deepens our understanding of voice persona modulation, offering valuable insights for advancing TTS technology. By bringing the acoustic underpinnings of vocal persona transitions to light, our findings aim to contribute to more expressive and engaging TTS applications, with broader implications for voice branding, assistive technology, and human-computer interaction.

Contents

1 Introduction

In interpersonal communication and social interaction, the human voice conveys rich information. Upon hearing someone speaking, listeners can extract nonverbal information embedded within speakers' vocal expressions such as gender, age, emotion and speaker identity, beyond decoding the literal linguistic messages (Zäske, Schweinberger, & Kawahara, 2010). Vocal persona refers to the identity or character perceived by others through an individual's voice, reflecting the intrinsic relationship between one's voice and the multifaceted dimensions of identity, emotions, and behavioral tendencies, as outlined by Tagg (2012). According to Pisanski, Cartei, McGettigan, Raine, and Reby (2016) and Noufi, May, and Berger (2023), individuals dynamically tailor their vocal personas to adapt to diverse social contexts and conversational intents. For example, a person employs an "authoritative" vocal persona to convey confidence and knowledge in a professional setting, while adopting a "friendly" voice to express warmth and openness during casual conversations with friends. This adaptability illustrates how individuals transition between different vocal personas to present their desired social presences and achieve their communicative intents across various contexts.

The ability to modulate vocal personas across social interactions, whether expressing friendliness in a social setting, projecting confidence in a professional scenario, or asserting dominance in a hierarchical situation, is crucial not only for human communication but also for synthesized speech. While current text-to-speech (TTS) systems have made significant advancements in achieving intelligibility in synthetic voices, a significant challenge persists in attaining a high degree of naturalness and expressivity. This gap in expressivity emphasizes the need to integrate adaptability into TTS systems, enabling them to dynamically respond to diverse social contexts and communication needs through vocal persona transitions. By incorporating this adaptability, synthesized speech can achieve greater expressiveness, authenticity, and user engagement.

While previous research has primarily focused on investigating emotional states within TTS expressivity, recent studies by F. Chen et al. (2004) and Moine and Obin (2020) have highlighted a broader spectrum of expressive speech, which extends beyond primary emotions to encompass nuanced social attitudes. Attitudes, as conceptualized within a bidimensional framework proposed by Aucouturier and Canonne (2017), encapsulate varying degrees of hostility/friendliness towards others and an individual's position within a social hierarchy, spanning from subordination to dominance. These attitudes include traits like friendliness, dominance, and distance. Understanding and representing these nuanced attitudes beyond fundamental emotional states are pivotal for enhancing the expressiveness and authenticity of synthesized speech (F. Chen et al. (2004). For instance, if you interact with your virtual assistant Siri to play music or ask for the weather forecast, a friendly synthetic voice would foster a positive interaction by creating a warm atmosphere. Similarly, if you seek information on complex topics or historical events from Siri, a synthesized voice with an authoritative attitude would enhance the credibility of responses and user trustworthiness. These examples demonstrate that incorporating nuanced social attitudes such as friendliness or professionalism, beyond basic emotions like happiness, into synthesized speech can better adapt to diverse social contexts and communicative needs, ultimately enhancing user engagement and TTS expressivity. Thus, there is a pressing need to bridge the gap in understanding and synthesizing more nuanced affective states, particularly social attitudes across diverse social settings, in expressive TTS systems.

Among these attitudes, a friendly attitude fosters warmth, trust, and mutual respect, essential for effective communication and relationship-building across various social contexts. For instance, customer service representatives employ friendly speech to express empathy and

understanding, thereby reassuring customers and enhancing satisfaction. Likewise, colleagues adopt a friendly tone to encourage collaboration and create a supportive environment in the workplace. These examples underscore the significance of friendly attitudes in facilitating effective human communication across various social settings. Moreover, the importance of friendliness extends to synthetic speech, where it plays a crucial role in expressiveness and engagement in voice interfaces such as voice assistants and conversational/virtual agents (Moine & Obin, 2020), including Siri, Alexa, and Google Assistant. For example, if Alexa can respond with a friendly attitude rather than an "average" tone when engaging in conversations about hobbies and plans with users at home, this friendly voice leaves an impression of chatting with a friend, thereby fostering a sense of connection and trust between the user and the virtual assistant. Embodying a friendly persona into these voice interfaces humanizes the interaction, creating more natural and conversational interactions, ultimately enhancing user engagement.

A friendly attitude in TTS systems is crucial for effective interactions with speech technologies. Despite this significance, there is a noticeable lack of emphasis on attitudinal speech synthesis, particularly within TTS systems aimed at conveying friendliness. To the best of our understanding, there is currently no research directly targeting synthesizing friendly speech. This gap likely exists because synthesizing expressive friendly speech requires a vast amount of data in a friendly attitude, yet expressive speech datasets are scarce and often limited to emotions (Moine & Obin, 2020). This greatly constrains the scope of research on expressive speech, posing challenges in directly synthesizing friendly speech. Therefore, an indirect approach to synthesizing a friendly voice is necessary.

While the direct synthesis of friendly speech remains challenging, the concept of vocal persona shifting offers valuable insights as an indirect approach to address this challenge. Vocal personas, inherently dynamic and adaptable, suggest the possibility of transitioning from one attitude to another, including shifting towards a friendly vocal persona. Such vocal persona shifting is not arbitrary but is closely linked to vocal parameters, such as pitch, intonation, and timber (Noufi et al., 2023; Tagg, 2012). According to Noufi et al. (2023), vocal personas are inherently equipped with a set of acoustic attributes aligned to specific contexts, allowing for rapid adjustment of vocal expressions in response to immediate situational change. Therefore, we can transit to different personas through a set of acoustic manipulations. Specifically, a friendly vocal persona is likely achieved by transitioning from another persona through acoustic modulation.

Existing studies that explore friendly speech mainly focus on examining acoustic differences between neutral speech and friendly speech in Chinese Mandarin. These studies suggest that the distinctions and connections in acoustic features between neutral and friendly vocal personas lie in pitch, duration, and energy (F. Chen et al., 2004; Li et al., 2004; Li & Wang, 2004). Building upon these findings, we can synthesize friendly speech from a neutral vocal persona through acoustic modulations in pitch, duration, and energy, while maintaining linguistic content and speaker identity.

This research focuses on synthesizing friendly speech from neutral speech by modifying key acoustic features—pitch, duration, and energy—using the FastSpeech 2 (FS2) framework, facilitating the perceptual transition of vocal personas from neutral to friendly in Mandarin Chinese TTS. Noufi et al. (2023) suggested that individuals adjust their vocal personas based on their orientation towards a specific language or culture, implying that acoustic modifications might be language-specific for a certain vocal persona. Given that previous studies on the acoustic analysis of friendly speech have primarily focused on Mandarin Chinese, Mandarin Chinese is chosen as the focal language for this study. Additionally, FS2 is selected due to its well-known speed, robustness, and high-quality speech output. Importantly, FS2 also pro-

vides controllability, allowing precise adjustments of pitch, duration, and energy independently through three distinct predictors (Ren et al., 2020). Leveraging the FS2 TTS system enables this research to manipulate acoustic features effectively.

By employing neutral speech data rather than friendly speech data for this synthesis, the study aims to address the challenge of limited expressive data in social attitudes, particularly in terms of friendliness. Furthermore, the integration of subtle social attitudes like friendliness into TTS synthesis aims to bridge the existing gap between the current capabilities of TTS systems and the demand for more nuanced representations of social attitudes in synthesized speech. Consequently, this research aims to contribute to the advancement of the expressiveness and user engagement of TTS systems by providing insights into the synthesis of friendly speech. It also explores its potential implications for enriching interactions with speech technologies.

1.1 Research Question and Hypothesis

This research seeks to systematically explore how modifications in key acoustic parameters—pitch, duration, and energy—can facilitate the perceptual transition of vocal personas from neutral to friendly in Mandarin Chinese TTS, employing the FastSpeech2 (FS2) framework. The main research question is formulated as:

> **How do specific manipulations of pitch, duration, and energy influence the perception of friendliness in synthesized Mandarin Chinese speech?**

Li and Wang (2004) and Li et al. (2004) suggested that pitch is the primary cue for perceiving friendliness in synthetic Mandarin Chinese, but optimal friendliness synthesis requires a combination of pitch, duration, and energy. Therefore, we propose to explore the individual and combined effects of these acoustic features on enhancing the perceived friendliness of synthesized speech. F. Chen et al. (2004) reported that friendly speech exhibits a higher mean pitch, shorter phone duration, and slightly elevated spectral energy compared to neutral speech. Based on these findings, we formulated two hypotheses:

- **Manipulating the mean pitch of a neutral voice upward will notably enhance the perception of friendliness.**
- **Incorporating pitch modulation with shorter phone duration and a slight increase in energy levels will produce a more pronounced perception of friendliness in synthesized speech compared to pitch modulation alone.**

1.2 Thesis outline

After outlining the context of the research and its rationale in this section, the rest of the thesis is structured as follows: Section 2 provides a comprehensive literature review presenting relevant research conducted in related topics. Section 3 details the methodology, including descriptions of the dataset, model, and procedure. Section 4 describes the obtained results, while Section 5 interprets these results and discusses potential explanations. Lastly, Section 6 concludes the thesis and presents recommended future work, highlighting its relevance in the field.

2 Literature Review

This section provides a comprehensive overview of prior research focusing on various aspects of synthesizing friendly speech, which covers the perception, acoustic analysis, and the synthesis of attitudinal speech, with a particular emphasis on friendly speech. Through an extensive analysis and critical synthesis of the literature, this review aims to establish the context for the current study on transitioning vocal personas from neutral to friendly through acoustic manipulations.

This section starts with a detailed description of the keywords used in the literature search, followed by the inclusion/exclusion criteria applied during the literature selection process. Subsequently, comprehensive summaries of the selected papers are presented, organized into subsections ranging from 2.1 to 2.7, each dedicated to addressing a specific topic.

The literature search was primarily conducted on Google Scholar, employing strategies such as keyword search and citation search. The chosen keywords were organized based on specific topics relevant to this study, as summarized in Table 1. Citation search involved reviewing the reference lists of identified papers to find additional relevant sources cited within them, thus identifying research on similar or related subjects.

Table 1: Topics and their corresponding keywords searched for literature

Topics	**Keywords**
Speech perception	vocal persona, speech perception, speaker perception
Attitudinal speech	attitudinal speech, voice and attitudes, acoustic analysis of attitudes
Friendly speech	friendly speech, friendly speech synthesis
TTS models	expressive TTS, expressive stylistic TTS model, controllable TTS
Acoustic manipulation	acoustic manipulation, paralinguistic manipulation, controllable TTS

Subsequently, a set of inclusion and exclusion criteria was applied to refine the selection process, ensuring the incorporation of relevant and up-to-date literature that aligns with the research objectives of this study.

1. Given the limited research on the acoustic analysis of friendly and neutral speech in Mandarin Chinese, this study selected relevant papers from various languages and attitudes, without restricting solely to Mandarin Chinese or friendly speech. The aim was to offer a comprehensive analysis of acoustic cues in attitudinal speech;

2. To prioritize the most influential works, I selected the top 20 articles based on the relevancy on Google Scholar;

3. Only English publications from peer-reviewed journals that focus on healthy participants were selected from the search results.

The remaining parts of the literature review are structured into 7 subsections based on subtopics related to this study. Subsection 2.1 introduces the concept of vocal persona and the proposed framework. The perception of speech, particularly in attitudinal speech, is described in subsection 2.2. Subsection 2.3 discusses the role of acoustic cues in attitudinal speech in general, while subsection 2.4 focuses on acoustic correlates in a friendly attitude, particularly

in Chinese Mandarin. In subsection 2.5, the prior work on friendly speech synthesis is covered. The models employed in expressive speech synthesis are discussed in subsection 2.6, to provide context for the chosen model in this study. Finally, a literature summary is provided in subsection 2.7, integrating insights from various subsections to identify research gaps and foster a comprehensive understanding of the research topic.

2.1 Vocal persona

The concept of vocal persona, as proposed by Tagg (2012), refers to the unique identity or character perceived by others through an individual's voice, including the speaker's gender, age, emotion, attitudes and behaviour. Noufi et al. (2023) further define it as a selected set of vocal expressions to adapt and react within a given communication setting, emphasizing the role of contexts in shaping vocal expressions and the fluid nature of vocal persona. As per Noufi et al. (2023), individuals typically shift between distinct vocal personas throughout the day, averaging 3.3 different personas daily, in order to adapt to social dynamics across diverse situations.

The vocal persona framework proposed by Noufi et al. (2023) elucidates how individuals choose vocal expressions to present vocal personas within contextualized communication, as illustrated in Figure 1. The process begins with one's intent in a specific context, leading to the selection of a persona from a set of predefined personas. Once the persona is chosen, corresponding paralinguistic attributes, such as pitch and speech rhythm, are selected to match the persona's characteristics and specific contexts. Subsequently, the resulting vocal expression is articulated and then refined based on feedback from intended audiences and the speaker's own assessment. The entire process is cyclical, involving adjustments of expressive attributes to better convey a desired persona, or switches of persona to adapt to different contexts.

To exemplify the process, let's consider a scenario where a speaker intends to convey a friendly attitude to a listener in an informal setting. In this predetermined context, a friendly vocal persona is selected, embodying traits like warmth, empathy, kindness, openness. The selection of this vocal persona entails paralinguistic traits including a light tone, a faster speaking rate and a larger pitch variation, which may necessitate adjustments to these acoustic attributes — pitch, intonation, and speech rate — to align with the role of a friendly persona. To better embody this friendly persona, continuous refinement through acoustic adjustments informed by audience feedback and self-perception may be necessary. However, in situations where there's a sudden shift in social contexts, such as unexpectedly encountering a business partner while conversing with a friend, the speaker may need to promptly transition from a friendly persona to an authoritative one and initiate a new cycle. This iterative process aims to convey the desired vocal persona and facilitate effective communication.

The authors outlined a detailed connection between context, acoustic attributes and vocal personas. Specifically, contextualized environments induce shifts in vocal personas. Moreover, adopting a specific persona influences the production of paralinguistic attributes, while adjusting acoustic features impacts the perception of vocal persona. However, the precise connection between each persona and its corresponding paralinguistic features remains unestablished, and the mechanism for transitioning from one vocal persona to another via adjustments in acoustic features is not investigated in Noufi et al. (2023)'s study. Therefore, further exploration in these areas is warranted.

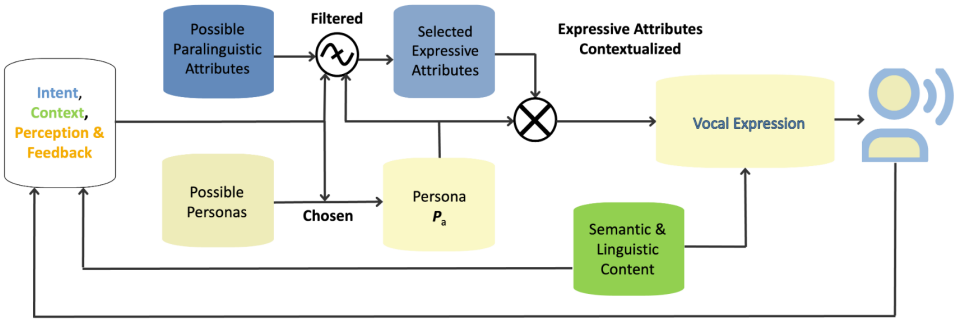


Figure 1: Flowchart depicting the process of persona-guided vocal expression (from Noufi et al. (2023)

2.2 Perception of attitudinal speech

Similar to visual cues, voices also carry a multitude of information about an individual's identity, emotions, attitudes, and behavioral positions (Tagg, 2012). Previous studies have demonstrated that various long-term and stable speaker traits can be perceived solely from voices, including gender (Huang, Duan, & Lyu, 2021; Schuller et al., 2010), age (Schuller et al., 2010), ethnicity (Omar & Pelecanos, 2010) and personality (Ivanov, Riccardi, Sporka, & Franc, 2011; McAleer, Todorov, & Belin, 2014; Polzehl, Moller, & Metze, 2011), even from a single word like 'hello' (McAleer et al., 2014).

In addition to encoding demographic information of speakers, voices also convey characteristics related to the speaker's short-term states, such as speech emotions, which refer to one's internal emotional states, and social attitudes, which refer to one's outward expression. Several studies have explored the intricate relationship between an individual's voice and their emotional states. Scherer and Scherer (2011) have reported a high accuracy in recognizing six primary emotions from audios, comparable to the performance based on visual cues, revealing the ability to perceive emotions from voices. Likewise, Schirmer and Adolphs (2017) have also suggested the capacity to perceive emotions conveyed by speakers through their voices.

While existing research has primarily focused on the relationship between voice and emotional expression, the perception of attitudes is equally crucial but remains underexplored (Moine & Obin, 2020). Noufi et al. (2023) explored the intricate interplay between internal emotional states and outward expression of attitudes, suggesting a complex relationship. They proposed that outward expression may intentionally or unintentionally disconnect from one's internal state depending on communicative contexts. Specifically, speakers often modulate or mask internal emotional states to present certain outward attitudes towards others, thereby controlling how their voices are perceived within certain contexts. For instance, a salesman may suppress anger to convey a friendly persona to customers in their work, projecting trustworthy and kind. This prioritization of context in communication underscores the crucial role of attitudes in shaping perceptions of voices.

Voice serves as a social tool for self-expression in everyday social interactions (Pisanski et al., 2016). However, there remains a notable gap in our comprehension of how voices convey social attitudes. This lack of exploration into attitudinal speech stems from their fuzzy nature in both perception and acoustic expression, as well as their heavy reliance on speech context, as highlighted by F. Chen et al. (2004).

2.3 Acoustic cues in attitudinal speech

Tagg (2012) proposed that specific acoustic parameters can convey a speaker's social attitudes, as well as personal or sociocultural identity, feelings, and emotions, highlighting the significant role of nonverbal paralinguistic features in expressing attitudes. Noufi et al. (2023) further emphasized the crucial role of acoustic features in expressing various social attitudes, claiming that paralinguistic attributes such as pitch and intonation are instrumental in conveying a speaker's vocal persona, including their attitudes.

Additionally, several studies have explored how distinct attitudinal speeches – such as authoritative, friendly, and dominant speech – are communicated through a set of acoustic parameters, including pitch, duration, energy, and phonation type. While some acoustic parameters have been consistently observed across attitudinal voices, others have been uniquely associated with specific attitudes. Research has also investigated this relationship across languages, revealing both universal and language-specific patterns. Table 2 summarizes these findings, illustrating how acoustic features can effectively convey different social attitudes across languages.

Pitch

It is widely recognized that pitch serves as the primary acoustic indicator conveying information about a speaker's traits, emotional states, and attitudes (Ponsot, Burred, Belin, & Aucouturier, 2018). For attitudinal speech, pitch appears to be an important acoustic correlate across various social attitudes. Authoritative voice is consistently associated with a lower mean pitch compared with a controlled voice, independent of languages. Sorokowski et al. (2019) elicited authoritative and controlled speech production from scientists working at various universities in Poland by prompting them to provide either expert opinions or directions. They found that the mean pitch in authoritative speech involving expert opinions is lower than in control speech by conducting an acoustic analysis. In addition to Polish, the Lachixío Zapotec language spoken in Mexico also displayed a lower F0 pattern during authoritative speech, as evidenced by a case study employing descriptive analysis conducted by Sicoli (2010).

Likewise, pitch is considered the most perceptually salient acoustic characteristic of the voice, influencing perceptions of dominance (Aung & Puts, 2020). Puts, Gaulin, and Verdolini (2006) investigated the correlation between pitch and dominance within a mating context in English. They manipulated pitch using computer software and subsequently conducted a dominance perception experiment. Their findings revealed that a lower-pitched voice tends to enhance the perception of dominance, particularly among males. Similar results were reported in Mandarin Chinese by Liu, Zhang, and Liang (2023), who extracted voices from characters of varying social status from a popular Chinese TV show. They found that dominant speech produced by dominant characters had lower pitch compared to subordinate characters. Consistent with these studies, Ponsot et al. (2018) manipulated pitch in recorded voices through a voice-processing algorithm and examined perceived dominance in French. They suggested that mean pitch is negatively related to one's perceived dominance, indicating that lower pitch correlates with higher dominance. Most studies have reported a negative relation between pitch and dominance in speech because a lower fundamental frequency (F0) is associated with conveying the impression of a larger body size, which is often linked to traits like dominance or aggressiveness (Geng, Gu, Johnson, & Erickson, 2020), based on the frequency code hypothesis proposed by Ohala (1984).

However, Salais et al. (2022) concluded that speakers raised their pitch in dominant speech

after analyzing the vocal production of four different attitudes in French, including dominant attitude. They accounted for this discordance by noting that previous studies compared dominance with neutral speech, rather than with other vocal attitudes, as their study did. Additionally, different language settings and culturally learned vocal associations may also contribute to this distinction. Similarly, Geng et al. (2020) elicited dominant and submissive speech from 33 native Mandarin speakers through role-play dialogue and reported that dominant voices having a higher F0 compared to submissive voices. The authors accounted for this opposite pattern simply by attributing it to differences in speakers or materials.

Pitch is also believed to be the most prominent acoustic cue for friendliness in speech (Li & Wang, 2004), with consistent patterns observed across languages. In English, Noble and Xu (2011) modified naturally produced neutral-emotion utterances in median pitch, pitch range, formant shift and voice quality, and subsequently conducted a perception test assessing happiness and friendliness employing these modified stimuli. They found that a higher mean pitch is exhibited in perceived friendly speech. Ranganath, Jurafsky, and McFarland (2013) also conducted an acoustic analysis of friendly speech in English, along with other attitudes, using a spoken corpus from speed-dates. They reported that women's friendly speech exhibited a higher maximum pitch. Similarly, higher pitch is attributed to friendly speech in Mandarin Chinese. In F. Chen et al. (2004), two male drama school students were recruited to record speech reflecting eight different attitudes, including friendly and neutral ones. Subsequently, a perception test was conducted to evaluate pairs of sentences whether they are perceptually distinct in terms of neutrality and friendliness. An acoustic analysis was then executed on the perceptually distinct sentences, with a focus on pitch, duration, and energy. The study revealed an increase in mean pitch for speech characterized by a friendly attitude. These findings align with cross-lingual literature on French (Salais et al., 2022) and Swedish (House, 2005) and can be explained by Ohala (1984)'s frequency code hypothesis, which posits that high acoustic frequency projects a smaller vocalizer and body size, conveying messages of submissiveness, non-threatening intent, a desire for the receiver's goodwill, and friendliness.

In general, the attitudinal speech discussed above highlights the importance of pitch in perceiving a speaker's attitudes and emphasizes a robust correlation between attitudes and pitch, regardless of the specific attitude or language being considered.

Duration/speaking rate

In addition to pitch, duration is another crucial acoustic attribute investigated in attitudinal speech research. Geng et al. (2020) proposed that dominant speech in Chinese Mandarin tends to exhibit shorter durations compared to submissive speech. Similarly, various studies have also highlighted this tendency for friendly speech in Chinese Mandarin. For example, Li and Wang (2004) observed that friendly speech tends to be faster than neutral speech. Likewise, F. Chen et al. (2004) found that friendly speech manifests shorter phone durations compared to neutral speech. Furthermore, Tang and Gu (2015), through an acoustic analysis of 13 attitudinal styles elicited from two university students, concluded that friendly speech is characterized by a faster speaking rate. These findings collectively emphasize the significance of duration in conveying different social attitudes, particularly in the context of friendly speech. Conversely, Sorokowski et al. (2019) observed no significant effect of duration in authoritative voice compared to controlled voice in Polish.

The exploration of duration in attitudinal speech primarily focuses on Mandarin Chinese, revealing a consistent pattern of shorter durations in both dominant and friendly speech. However, the investigation of the relationship between duration and social attitudes in other lan-

guages remains limited, except for the study conducted in Polish by Sorokowski et al. (2019).

Energy

Energy serves as another acoustic cue in influencing attitudinal speech, yet it has received relatively limited attention. Geng et al. (2020) observed significant variations in both the mean and range of intensity between dominant and submissive attitudes in Chinese Mandarin. Specifically, utterances associated with a dominant attitude displayed higher mean intensity and a broader intensity range. Similarly, F. Chen et al. (2004) noted varying degrees of increase or decrease in energy levels across different frequency ranges in friendly speech compared to neutral speech in Mandarin Chinese. However, no other studies have examined energy in authoritative speech or in languages other than Mandarin.

Other acoustic parameters

Other acoustic parameters examined in attitudinal speech include pitch variation, formants, and phonation types. Specifically, a wider F0 range is reported to be associated with dominant attitudes when contrasted with submissive speech in Mandarin (Geng et al., 2020). Similarly, English and Dutch tend to demonstrate a larger pitch variance in friendly speech (A. Chen, Rietveld, & Gussenhoven, 2001; Noble & Xu, 2011; Ranganath et al., 2013), while friendly speech in Mandarin is suggested to exhibit low F0 variation (Tang & Gu, 2015). Phonation type consistently reveals discernible patterns in speech linked to various attitudes. Creak serves as a communicative tool for asserting authority in interactions (Callier, 2013; Hildebrand-Edgar, 2016), commonly observed in authoritative and dominant speech across languages (English: Yuasa (2010), Meier (n.d.); Chinese Mandarin: Liu et al. (2023), except for Lachix´ıo Zapotec language where authoritative speech is associated with a breathy voice (Sicoli, 2010). Conversely, breathy voice is reported in friendly speech, as observed in English (Noble & Xu, 2011). In terms of formants, authoritative speech tends to display low formant frequencies in comparison to controlled voice, as evidenced in Polish (Sorokowski et al., 2019). However, no significant differences in formants were observed between friendly speech and neutral speech in English (Noble & Xu, 2011).

2.4 Acoustic cues in friendly speech

The research focusing on paralinguistic study of friendly speech has been conducted in English (A. Chen et al., 2001; Noble & Xu, 2011; Ranganath et al., 2013), Chinese Mandarin (F. Chen et al., 2004; Li et al., 2004; Li & Wang, 2004), Dutch (A. Chen et al., 2001), French (Salais et al., 2022) and Swedish (House, 2005). As previously mentioned, pitch is a universal acoustic attribute in attitudinal speech, including a friendly attitude. Specifically, friendly attitude is always related with high mean pitch in Chinese Mandarin, English, French and Swedish (Mandarin: F. Chen et al. (2004); English: Noble and Xu (2011), Ranganath et al. (2013). While pitch appears to exhibit consistent patterns cross-lingually, variations were observed in the examination of other acoustic attributes in those languages. This included the investigation of different acoustic features or the observation of distinct acoustic patterns. Studies on friendly English and Dutch speech have primarily focused on examining acoustic parameters such as pitch variation and phonation type. A. Chen et al. (2001) revealed that both Dutch and English friendly voices demonstrated a larger pitch range, consistent with the findings in studies conducted by Ranganath et al. (2013) and Noble and Xu (2011). Additionally, Noble and Xu

Table 2: Acoustic cues in attitudinal speech by attitude and language

Type of attitudes	Language	Paper	Acoustic attributes					
			F0	Formant	Pitch variation	Phonation type	Duration	Energy
authoritative	Polish	Sorokowski et al. (2019)	Lower mean F0	Lower Formant position			No effect	
authoritative	Lachixío zapotec	Sicoli (2010)	Lower pitch			Breathy voice		
authoritative	English	Yuasa (2010)				Creaky voice		
dominant	Chinese mandarin	Liu et al. (2023)	Lower pitch			Modal/creaky voice		
dominant	English	Puts et al. (2006)	Lower pitch					
dominant	French	Ponsot et al. (2018)	Lower pitch		Pitch contour: pitch fall gradually			
dominant	Chinese Mandarin	Geng et al. (2020)	higher mean F0 (male only)		Wider F0 range		Shorter duration	higher mean intensity and wider range of intensity
dominant	French	Salais et al. (2022)	higher pitch					
friendly	Chinese Mandarin	Tang and Gu (2015)			low F0 variation		fast speaking rate	
friendly	Chinese Mandarin	F. Chen et al. (2004)	higher mean pitch				shorter phone duration	different patterns for energy level at different frequency ranges
friendly	English	Noble and Xu (2011)	higher mean pitch (200 Hz)	1.0 formant ratio	bigger pitch range (3 log scaled)	breathy voice		
friendly	English	Ranganath et al. (2013)	higher max pitch		greater pitch variance			
friendly	Chinese Mandarin	Li and Wang (2004)					faster	
friendly	French	Salais et al. (2022)	higher pitch					dynamic
friendly	Swedish	House (2005)	higher pitch					
friendly	English & Dutch	A. Chen et al. (2001)			bigger pitch range			

(2011) also identified a correlation between friendliness and breathy voice. In contrast to the acoustic analysis conducted in English, the study of Chinese Mandarin focused on different acoustic parameters, namely duration and spectrum energy. Research suggests that friendly speech in Mandarin is characterized by shorter phone duration/faster speaking rate, as well as higher mean pitch and slightly increased spectrum energy (Li & Wang, 2004; Tang & Gu, 2015).

In addition to language-specific patterns, the significance of each acoustic parameter in perceiving friendliness has been explored in Mandarin Chinese. Li and Wang (2004) initially conducted a perception test on the IBM-CASS expressive speech corpus to select perceptually distinct neutral and friendly sentence pairs. They subsequently employed the Psola synthesizer in Praat to synthesize friendly audio stimuli by adjusting either single or combinations of acoustic parameters (pitch and duration) from corresponding neutral voices. Another perceptual test was then conducted to examine which acoustic feature or feature combinations are most important in synthesizing friendly speech. The findings indicate that solely adjusting duration does not lead to expressive friendly speech; however, modulating pitch does. In 2004, Li et al. (2004) employed a similar method and further investigated the impact of energy on perceptions

of friendliness. They synthesized friendly speech using IBM's formant TTS system under various conditions, including manipulations of single acoustic features, pitch, duration, and energy, or different combinations thereof. Participants were then recruited for a perception test to listen to these synthetic audio samples and rate the perceived friendliness following different acoustic manipulations. The results demonstrated that adjusting pitch led to significant differences in participants' perceptions of friendliness, whereas modulating duration or energy alone did not significantly impact friendliness perception. Moreover, the optimal adjustment for friendliness was found to be the combination of pitch, phone duration, and spectral energy.

2.5 Friendly speech synthesis

According to Moine and Obin (2020), a friendly attitude embodies pleasant, benevolent characteristics and care towards others. This attitude is essential for effective communication and relationship-building, as well as in enhancing expressivity in TTS systems. Despite the importance of friendly speech, to the best of our knowledge, no research has directly targeted the synthesis of expressive voices conveying a friendly attitude. Instead, two studies conducted by Li and Wang (2004) and Li et al. (2004), which have been previously discussed, indirectly synthesized friendly speech from neutral speech in Chinese Mandarin using neutral speech data. Li and Wang (2004) synthesized friendly stimuli by adjusting the acoustic parameters of neutral voices using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat. PSOLA is an algorithm that adjusts the pitch and duration of speech segments, allowing for the manipulation of pitch, duration, or a combination of both. Li et al. (2004) synthesized friendly speech stimuli using IBM's formant TTS system. This system generates speech by manipulating formant frequencies, allowing precise control over individual acoustic features, including duration, pitch, energy, and combined acoustic modulations.

While the primary aim of these two studies was to investigate which acoustic parameters or combinations contribute most to the perception of friendliness, these studies also shed light on an indirect approach to synthesizing friendly speech. Through the modulation of various acoustic features, both studies underscored that the synthesis of friendly speech could be achieved by adjusting distinct acoustic parameters from neutral speech. Furthermore, they suggested that the optimal friendliness is achieved by jointly manipulating three key acoustic features: pitch, duration, and energy, with pitch modulation alone significantly enhancing speech friendliness. However, it's noteworthy that both studies utilized outdated TTS techniques, such as the Psola synthesizer in Praat and BM's formant TTS system. These older systems may fall short in accurately manipulating acoustics and producing high-quality expressive friendly speech. Additionally, the methodology employed in these two studies may not be entirely valid. The perceptual tests involved participants rating how friendly the stimuli sounded using MOS, which did not include decoys. The absence of decoy answers may signal to participants that the study specifically concerns friendly speech, potentially priming them to perceive the stimuli in a friendly manner. This can influence their responses and compromise the validity of the results, as participants' perceptions may be biased by the study's focus on friendliness. Therefore, a new study with a more valid method design and a state-of-the-art TTS system is necessary to validate the effectiveness of such acoustic manipulation on the perceptual transition of vocal persona.

2.6 Expressive and controllable TTS

TTS systems have significantly advanced over time, achieving remarkable intelligibility and near-human naturalness in synthetic voices through deep neural network advancements. However, neural network based TTS has been criticized in terms of the lack of expressiveness (Lee, Park, & Kim, 2021; Wang et al., 2018). That is due to the fact that neural TTS models learn an average prosodic distribution from the training data using mean loss, leading to the generation of less expressive voices (Lee et al., 2021; Wang et al., 2018).

To achieve expressive voice synthesis, researchers have developed expressive TTS systems like Tacotron 2 with global style tokens (GSTs). These systems aim to learn and model various styles, such as intention and emotion (Wang et al., 2018). By incorporating prosody into reference embedding, GSTs generate soft and interpretable tokens representing various prosodic features such as pitch, speaking rate, and energy. These prosodic tokens enable precise control over styles through either style control or style transfer tasks. For instance, style control allows users to adjust specific parameters such as pitch and speaking rate to convey different emotions or attitudes, while style transfer enables the system to adopt the prosodic characteristics of a particular style or speaker. This capability significantly enhances expressive synthesis by generating more natural, varied, and contextually appropriate speech, effectively conveying different emotions and intentions (Wang et al., 2018). Nevertheless, autoregressive expressive TTS systems yield to slow training and inference speed because of its autoregressive architecture that the prediction of each frame is conditioned on all previous time steps (Lee et al., 2021). Additionally, it also has a limitation of robustness, that is, word skipping and repeating issues due to inaccurate attention alignments during inference.

To address issues, non-autoregressive TTS systems were proposed. Among these methods, FastSpeech 2 (FS2), a state-of-the-art non-autoregressive TTS system introduced by Ren et al. (2020), gained popularity. FS2 utilizes feed-forward Transformer blocks, consisting of stacked self-attention layers (Vaswani et al., 2017) and 1D-convolutions in both its encoders and decoders (Ren et al., 2020). This architecture enables parallel decoding, resulting in a notably faster inference speed. In addition, the substitution of the attention mechanism into a length regulator addresses the robustness issue, leaving almost no word skipping and repeating as in autoregressive TTS. Furthermore, FS2 employs a variance adaptor, including separate duration, energy and pitch predictors to provide variation information to predict variant speech, thereby addressing the one-to-many mapping problem in the TTS field and enhancing the expressiveness of synthetic voices. The integration of the GST approach with FastSpeech 2 further extends its capabilities. As previously mentioned, GSTs enable the model to learn and reproduce different styles, allowing FS2 to adapt its output to diverse emotions, attitudes, speaking styles, and beyond. This versatility greatly enriches the expressiveness of the synthesized speech. However, this approach requires a substantial amount of expressive speech data to effectively train the GSTs. Yet, there is still a notable scarcity in the availability of such expressive speech datasets.

In addition to expressivity, TTS models have faced criticism for the lack of controllability (Lee et al., 2021; Wang et al., 2018), which refers to the users' ability to control output properties with adjustable parameters alongside the textual input in controllable TTS systems (Henter, Lorenzo-Trueba, Wang, & Yamagishi, 2017). This capability offers greater flexibility and customization options. Historically, different approaches to controllable TTS have evolved over time, each with its own strengths and limitations. According to Tits, El Haddad, and Dutoit (2021), formant synthesizers allowed control over numerous parameters but produced unnatural voices. Concatenative TTS subsequently emerged, producing more natural voices but

with limited controllability. Statistical Parametric Speech Synthesis (SPSS), such as Hidden Markov Models (HMM), offers a balance between naturalness and controllability. It offered fair naturalness while providing a certain level of control over parameters. Later, Deep Neural Networks (DNN) revolutionized TTS by achieving high naturalness. Nevertheless, the demand for extensive annotated data, particularly for style or emotion data, necessitated a shift to unsupervised strategies (Tits et al., 2021). Tacotron2 with Global Style Tokens (GSTs) has emerged as a promising approach, enabling control over speaker identity, speaking style, and prosodic information (Wang et al., 2018). However, due to its autoregressive nature and the absence of explicit alignments between text and speech, directly controlling voice speed and prosody poses difficulties (Ren et al., 2019).

FastSpeech tackles the issue of voice speed regulation through a length regulator based on the formula:

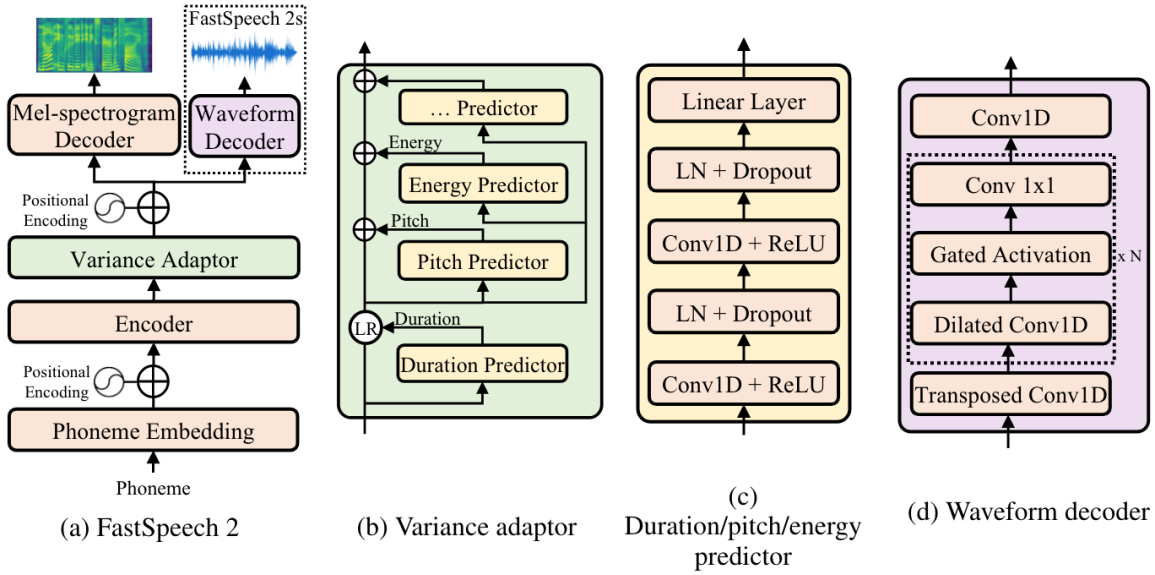
$$Hmel = LR(Hpho, D, \alpha) \tag{1}$$

Specifically, it computes the expanded sequence *Hmel*, representing phoneme duration, by multiplying the hidden states of the phoneme sequence *Hpho* with the phoneme duration D predicted by the duration predictor. The accurate prediction of phoneme durations crucially depends on precise phoneme-level alignments, which are achieved by the external aligner Montreal Forced Aligner (MFA) (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) within the FastSpeech framework. Furthermore, the hyperparameter α further modulates phoneme duration. If α is larger than 1, the duration sequence gets expanded, leading to longer phoneme durations and slower speaking rates. Conversely, a value of α less than 1 yields shorter phoneme durations and faster speaking rates, while α equals 1 produces normal speed. Additionally, by manipulating the duration of space characters, FastSpeech can regulate breaks between words, further adjusting prosodic aspects of synthesized speech (Ren et al., 2019).

In addition to duration modulation, FS2 allows for precise and flexible control over pitch and energy through its variance adapter, which includes a duration predictor, a pitch predictor, and an energy predictor. As illustrated in Figure 2, the duration, pitch, and energy information extracted from these predictors are integrated into a phoneme hidden sequence obtained from the encoder. This adapted hidden sequence, enriched with duration, pitch, and energy details, is then inputted into the decoder to generate mel-spectrograms. The introduction of variance information including duration, pitch and energy, provides diverse speech outputs from the same text input, thereby addressing the one-to-many mapping problem in TTS. As a byproduct, it enhances the controllability of synthesized speech, allowing for manual adjustment of pitch, duration, and energy (volume) in the synthesized audio through the hyperparameter α (Ren et al., 2020), similar to the length regulator in FastSpeech.

2.7 Summary

The literature review elucidates the pivotal role of acoustic features—particularly pitch, duration, and energy —in modulating vocal personas and influencing listener perceptions in the domain of speech synthesis. Studies highlight pitch as a dominant factor in conveying friendliness, suggesting its potential as a primary manipulator in voice persona transitions. Despite this foundational understanding, there remains a gap in applying these insights to Mandarin Chinese TTS systems, specifically in achieving nuanced persona transitions such as from neutral to friendly. Moreover, while individual impacts of these acoustic parameters are well-documented, their combined effects on listener perception in TTS contexts warrant further exploration. Therefore, this research seeks to bridge these gaps by investigating how the manipulation of these key acoustic features can facilitate a perceptual transition from a neutral to



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor (d) Waveform decoder

Figure 2: The overall architecture for FastSpeech 2 and 2s (from Ren et al. (2020)

a friendly vocal persona in Mandarin Chinese TTS systems using FS2—an expressive and controllable TTS system. By focusing on pitch, duration, and energy, this study aims to formulate an understanding of their interplay and overall impact on vocal persona transitions, informing more expressive and contextually adaptive TTS technologies. Consequently, the hypotheses emerge from these insights, positing that specific, controlled manipulations of these acoustic parameters can significantly enhance the perception of friendliness in synthesized speech. Specifically, simply elevating the average pitch of a neutral voice is anticipated to notably impact the perception of friendliness, while coupling this adjustment with shorter phone duration and a slight increase in energy levels is projected to further enhance overall friendliness (F. Chen et al., 2004; Li et al., 2004).

This study proposes integrating social attitudes into TTS systems to enrich the synthesized speech with nuanced attitudes, thereby enhancing expressivity and user engagement. Additionally, it aims to complement the vocal persona framework by elucidating the intricate connection between vocal personas and their associated paralinguistic features. Specifically, it focuses on the mechanism for transitioning between neutral and friendly vocal personas through adjustments in acoustic features. By implementing acoustic manipulation within the state-of-the-art TTS system FS2, this approach addresses methodological challenges in existing research on friendly speech synthesis and aims to validate the effectiveness of such manipulations in the perceptual transition from neutral to friendly vocal personas. The methodology used to address this gap is described in the following section.

3 Methodology

In this section, the methodology used in this study is described. First, the dataset utilized is introduced in subsection 3.1, followed by the model employed for training and synthesis, namely, FastSpeech 2, in subsection 3.2. Subsection 3.3 presents the details about model training, including MFA alignment and training parameters, while subsection refinference details the stimuli design used for synthesis and model inference. Subsequently, pilot studies are introduced in subsection 3.5, followed by a discussion on larger-scale listening tests, providing details on participant selection, procedure, test design, and measurements in subsection 3.6 Lastly, this section concludes with statistical analysis in subsection 3.7 and the ethical considerations inherent in this research in subsection 3.8. The code implementation and relevant procedures for this study are detailed in the GitHub repository[1].

3.1 Dataset - AISHELL3

As noted in Section 1, databases of expressive speech, particularly regarding attitudes, are scarce. Therefore, this study utilized existing neutral speech data in Mandarin Chinese and employed acoustic manipulation to transition it into friendly speech. Initially, three datasets were considered: CSS10 (Park & Mulc, 2019), Common Voice (Ardila et al., 2019), and AISHELL3 (Shi, Bu, Xu, Zhang, & Li, 2020). CSS10 comprises single-speaker speech datasets for multiple languages, including Chinese Mandarin data recorded by a native speaker. However, since the recordings were based on fiction books that may involve emotions and styles, they were excluded from this study. Common Voice, a large multilingual crowdsourced corpus, contains extensive Chinese Mandarin speech data. However, its crowdsourced nature introduces noise, making it unsuitable for TTS tasks that require high-quality speech data. The dataset chosen for this study was sourced from AISHELL 3[2], an open-source high-fidelity Mandarin speech corpus, comprising approximately 85 hours of emotional neutral speech data spoken by 218 native Mandarin speakers, totaling 88,035 utterances. This dataset encompasses a diverse range of speakers, including different genders, age groups, and native accents. In addition to audio recordings, transcripts in Chinese characters and pinyin are provided. These transcripts were annotated and verified by professionals, resulting in over 98% accuracy in word and tone transcription. The high-quality neutral Mandarin data in AISHELL 3 is well-suited for this study.

3.2 Model

To synthesize friendly speech, we initially considered using expressive TTS systems with style tokens, such as Tacotron 2 with GST. Style tokens capture and represent a friendly tone from the training data, thereby conveying a friendly attitude in the synthesized speech. However, finding expressive speech corpora specifically with a friendly attitude proved challenging. Therefore, we decided to generate expressive friendly speech indirectly from neutral speech data through acoustic manipulation. FS2, as discussed in subsection 2.6, is a non-autoregressive, expressive, and controllable TTS system that offers speed, robustness, and controllability. It allows for manual acoustic manipulation through separate predictors, making it well-suited for our needs. This study utilized a PyTorch implementation of FS2 [3] for both training and inference.

[1]The GitHub repository to the demo: `https://github.com/Chenyi063/friendly_speech_synthesis`

[2]Dataset can be found at: `https://www.aishelltech.com/aishell_3`

[3]PyTorch implementation of FastSpeech 2: `https://github.com/ming024/FastSpeech2`

3.3 Model training

In the FastSpeech2 GitHub repository, there is a pretrained model trained on the AISHELL-3 dataset. However, when we synthesized sentences using this pretrained checkpoint, the resulting audio quality was suboptimal, with noticeable clicks present. Importantly, the original AISHELL-3 dataset does not contain any click sounds, indicating that these clicks in the synthetic audio are likely introduced during the resampling process. Additionally, there was a bug in the implementation where pitch control was mistakenly applied in place of energy control, leading to inaccurate pitch and energy controls during inference. Due to these issues, we decided to train our own model on the AISHELL-3 dataset to achieve better audio quality and precise control over acoustic features.

3.3.1 Alignment

Before training the FS2 acoustic model, Montreal Forced Alignment (MFA) (McAuliffe et al., 2017) was used to obtain the alignments between the utterances and the phoneme sequences, in order to enhance alignment accuracy (Ren et al., 2020). MFA necessitates a paired text-audio corpus for training, so our initial step involved creating such a corpus along with necessary preprocessing. The AISHELL3 dataset features exceptional audio quality with minimal noise, thereby necessitating minimal preprocessing. However, to optimize computational resources without compromising speech quality, we resampled the audio files from a 44100 sampling rate to 22050. For text transcription, a Python script was used to preprocess a TXT file containing transcriptions of the entire dataset to extract transcripts for each audio file. These transcripts were then stored in LAB files, with each LAB file sharing the same name as its corresponding audio file. Finally, each LAB file and its corresponding WAV file were stored together in the same directory, forming a corpus.

Afterward, validation of the data directory structure was conducted to ensure compliance with the specific structures required by MFA. Since the AISHELL dataset is a multi-speaker dataset, it was structured according to the multi-speaker format requirement, as illustrated in Figure 3. This step yielded general information about the dataset and generated a TXT file containing the Out-Of-Vocabulary (OOV) words. Excessive OOV words can potentially compromise the quality of the TTS model; thus, their pronunciations were predicted using a pretrained Mandarin grapheme-to-phoneme (g2p) model and subsequently integrated into the pronunciation dictionary manually.

Subsequently, the aligner aligned orthographic transcription in LAB files and corresponding phonetic sequences in WAV files using a pretrained Mandarin acoustic model and the updated dictionary. This alignment process produced TextGrid files; however, upon examination of the TextGrid files alongside the corresponding WAV files in Praat, it was evident that the alignment of sounds and phone annotations was inaccurate, with the silence phone “sil” annotated for each phone. To address this issue, an acoustic model specifically tailored to this dataset was trained. The aforementioned steps were then repeated, resulting in new TextGrid files with improved alignments.

3.3.2 Training

Upon completion of the alignment process, the generated TextGrid files and corpus underwent preprocessing at the phoneme level to prepare for model training. Pinyin, the widely used romanization system for Standard Chinese, is used in the dictionary. However, as the Pinyin system comprising letters and tone combinations differs from English, it is essential to define

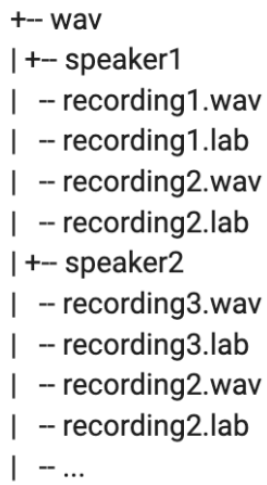


Figure 3: Multi-speaker corpus structure for MFA

this set of symbols used in text input to the model. Fortunately, the FS2 repository already includes Pinyin symbols, eliminating the need for additional steps in this regard. Following the preprocessing step, the acoustic model was trained on an Nvidia A100 GPU on Hábrók, which is a high-performance computing (HPC) cluster at the University of Groningen. The model underwent training for 600,000 steps using optimizer parameters as shown in Table 3. This number of training steps aligns with that utilized in the pre-trained Chinese model available in the FS2 repository, which serves as a reference. The entire training process lasted around 54 hours.

Table 3: Optimizer Parameters and Corresponding Values of the Trained Model

Parameter	**Value**
batch_size	16
betas	[0.9, 0.98]
eps	0.000000001
weight_decay	0.0
grad_clip_thresh	1.0
grad_acc_step	1
warm_up_step	4000
anneal_steps	[300000, 400000, 500000]
anneal_rate	0.3

3.4 Model inference

In the model inference phase of this study, stimuli for synthesis were initially designed and then synthesized with controllability using FS2. The design process of the stimuli, as well as the implementation of the inference, are detailed below.

3.4.1 Stimuli design

The sentences used for synthesis in this study were drawn from the research conducted by Moine and Obin (2020). Their study encompassed 100 sentences representing four distinct attitudes: dominant, friendly, seductive, and distant. These sentences were designed to be as neutral as possible and devoid of expressive content to minimize biases. Moreover, they were designed to be short with simple syntactic structures to reduce the prosodic variability. Furthermore, they were carefully crafted to ensure interpretability within each social attitude. These design choices are in line with the study's aim of avoiding semantically evoking participants' expressive status in listening tests, focusing on testing whether friendliness can be perceived based on paralinguistic cues rather than linguistic information. Consequently, 30 sentences were selected from Moine and Obin (2020)'s study and translated into Mandarin Chinese. The complete set of sentences used in this study is provided in Appendix A, including both Chinese sentences and their English translations.

3.4.2 Inference

During inference, each test sentence was synthesized under three conditions: Neutral (N), Pitch-modified (P), and Pitch-Duration-Energy modified (PED). The N condition involves no acoustic control, resulting in a neutral voice. In the P condition, pitch is modulated using the hyperparameter α, which denotes the manipulation rate. This rate indicates the factor by which the feature value is multiplied to achieve the desired modulation. For sentences under the PED condition, pitch, duration, and energy values are all manipulated independently with hyperparameters. Previous studies by Li and Wang (2004) and Li et al. (2004) also indirectly synthesized friendly speech from neutral speech by manipulating acoustic features pitch, duration, and energy. However, specific manipulation values were not provided in their studies. Alternatively, we derived specific hyperparameter values based on the acoustic analysis conducted by F. Chen et al. (2004). F. Chen et al. (2004) investigated acoustic differences on pitch, duration and spectral energy between naturally recorded neutral and friendly speech, reporting an increase in mean pitch, longer duration/faster speaking rate, and slightly higher frequency energy in friendly speech compared with neutral speech. However, acoustic differences between these two attitudes were distributed across different phonetic categories or tone groups. To ensure simplicity and controllability over mean values in FS2, we calculated these differences using average values, resulting in mean ratio values of 1.2475 for pitch, 0.9003 for duration, and 1.00017 for spectral energy. Consequently, the hyperparameter value for pitch was set to 1.2475, indicating that mean pitch values in the neutral condition are multiplied by this factor to achieve increased pitch. Similarly, hyperparameters for duration and energy control were set to 0.9003 and 1.00017, respectively. Additionally, a specific voice was selected from a range of voices the model was trained on, given that it's a multi-speaker model. Since the acoustic analysis comparing neutral and friendly speech was conducted using speech data from two male speakers in the reference study by F. Chen et al. (2004), the selected speaker voice was also set to a male voice to mitigate potential confounding factors. To ensure higher synthetic quality, the male voice with the most speech data among males in the dataset was selected.

After synthesizing audios, it was found that controlling pitch and energy in synthetic sentences presented a considerable challenge. Specifically, applying a specific control factor in pitch or energy did not result in a proportional change. For example, adjusting the pitch by a factor of 1.2475 did not result in sentences with 1.2475 times the F0 value (the objective measurements of pitch, measured in Hertz (Hz)) compared to those without pitch control. Fur-

ther investigation revealed that this discrepancy stemmed from the standardization of pitch and energy values during preprocessing. Applying control factors to the standardized pitch and energy led to a non-linear relationship between the original and modulated values. To resolve this issue, we initially identified the standardization formulas applied to pitch and energy during preprocessing:

$$value\_standardization = (value - mean)/std \tag{2}$$

Based on the formula, we obtained raw pitch and energy values from the standardized ones:

$$value = value\_standardization * std + mean \tag{3}$$

Subsequently, we applied control values to these raw pitch and energy values and then re-standardized these new raw values to derive new standardized values using the formula:

$$value\_standardization\_new = (value\_standardization * std + mean) * control - mean/std \tag{4}$$

These modifications enabled the effective control of the pitch and energy during inference.

After De-standardization of pitch and energy control, the output pitch and energy values still exhibited deviations from the expected values due to the rounding process. Ren et al. (2019) detailed the operation of the length regulator in FastSpeech with examples, shedding light on this discrepancy. The example involves the hidden states of the phoneme sequence *Hpho = [h1, h2, h3, h4]* and the corresponding phoneme duration sequence *D = [2, 2, 3, 1]*. Setting hyperparameter α = *1.3* results in duration sequences becoming *D*α*=1.3D = [2.6, 2.6, 3.9, 1.3]* $\approx$ *[3, 3, 4, 1]*, and the expanded sequences become *[h1, h1, h1, h2, h2, h2, h3, h3, h3, h3, h4]*. The symbol $\approx$ is utilized to denote the approximation to integers for control, contributing to the discrepancy between desired and realized pitch and energy values. To ensure precise control over pitch, duration, and energy—specifically, 1.2475, 0.9003, and 1.00017, respectively—we manually checked all the mean pitch, energy and duration values in each synthesized sentence. If the differences between the desired and actual values exceed ±5% for any acoustic feature, we adjust hyperparameter values α accordingly to approximate the desired ratio. Otherwise, we allow slight differences to persist due to the inherent challenge of achieving exact ratios for all acoustic parameters in a sentence. A detailed overview of control values for conditions P and PED in each sentence is provided in Appendix A, while stimuli synthesized for the study are referenced in Appendix D.

3.5 Pilot study

The success of nuanced attitude transitions from neutral to friendly lies in the precise control of acoustic parameters, particularly the specific values of hyperparameters for pitch, duration and energy control. To evaluate the effectiveness of these controls in perceptually transitioning vocal personas, a pilot study was conducted involving three native Mandarin speakers, including two males and one female. During the pilot listening test, each participant was presented with 10 sentences in random order, each accompanied by three audio stimuli under different conditions: N, P, and PED. Participants were then asked to select the perceived attitude from four predefined options: friendly, neutral, distant, and authoritative, for each stimulus. The results of the pilot study revealed that participants struggled to correctly identify attitudes from synthetic audio, particularly a friendly attitude, with average accuracy rates of 70%, 20%, and 17% for N, P, and PED conditions, respectively (see Table 4). Furthermore, participants had difficulty distinguishing among the four attitudes, as their choices were spread across all options without any consistent pattern.

The challenges in perceiving friendly attitudes in this pilot study contrast with the results reported in similar studies mentioned in subsection 2.5. This discrepancy may arise from the presence of additional contextual cues in those previous studies, rather than relying solely on acoustic cues as in the pilot study. As mentioned in subsection 2.5, the methodological setups in those studies induced priming effects, prompting participants to perceive the stimuli in a friendly manner and thereby introducing additional contextual cues alongside acoustic cues. Studies (Barrett, Mesquita, & Gendron, 2011; Kim & Kim, n.d.) suggested that listeners typically rely on multiple cues from different modalities simultaneously to perceive speech, such as auditory cues, visual cues, context, and speaker identity. Therefore, the inclusion of extra cues in addition to acoustic cues may enhance the perception of friendliness. In order to examine whether the difficulty in perceiving friendliness arises from the insufficiency of acoustic cues alone or from the unsuccessful acoustic manipulations, we introduced another version of the listening test. Acknowledging the priming effect of the MOS evaluation matrix, we intentionally employed MOS in the version 2 of the listening test to provide participants with additional contextual cues. This aimed to investigate if the introduction of contextual cues results in a stronger perception of friendliness from the same synthetic stimuli.

Following this version, we conducted another pilot study with three female participants, adhering to the same procedure. Participants were asked to rate the friendliness of each synthesized stimuli on a scale of 1 to 5 and then scores were averaged among these three participants to obtain MOS. The MOS serves as an indicator of the degree of friendliness perceived by participants, with a higher MOS indicating a stronger perception of friendliness. The results exhibited differences in MOS scores across three conditions, with average scores of 1.6, 1.9, 2.1 for conditions N, P and PED respectively (refer to Table 5). These findings suggested potential differences in the perception of friendliness under three conditions and the potential enhancement of friendliness perception following the introduction of contextual cues alongside acoustic cues. To further investigate, we proceeded to conduct larger-scale listening tests with both versions involving more participants.

Table 4: Individual and average accuracy in pilot study Version 1

Participants	Gender	Tech_back	Accuracy_N	Accuracy_P	Accuracy_PED
1	M	N	0.9	0.4	0.2
2	M	Y	0.5	0.0	0.2
3	F	Y	0.7	0.2	0.1
Average			0.7	0.2	0.17

Table 5: Individual and average MOS in pilot study Version 2

Participants	Gender	Tech_back	MOS_N	MOS_P	MOS_PED
1	F	N	1	1.0	1.3
2	F	Y	2.4	2.5	3.1
3	F	Y	1.3	2.1	1.9
Average			1.6	1.9	2.1

3.6 Listening test

Following pilot studies, two versions of larger-scale online listening tests were conducted to investigate the effects of acoustic cues alone and combined acoustic and contextual cues on the perception of friendliness, while also exploring differences in perceiving friendliness across three distinct conditions. The results of these two versions were compared to analyze their respective impacts. In each version, two hypotheses were examined: the effect of pitch modulation alone and the effect of combined pitch, duration, and energy modulations on the perception of friendliness. Both listening tests followed identical procedures for participant selection and testing but differed in their questions and evaluations.

3.6.1 Participants

The participants for this study were primarily recruited from personal networks, such as family, friends, and fellow students in the MSc voice technology program. Additionally, invitations to participate in the listening tests were distributed via social media platforms to recruit more participants beyond the personal network. In total, 46 native Mandarin speakers with normal hearing capacity participated in the test, with 24 individuals in version 1 of the listening test and 22 in version 2. Among these participants, the distribution of gender and background in each listening test was not balanced. Specifically, the number of females exceeded that of male participants, and participants without a background in speech technology outnumbered those with such a background. Details regarding the participants are provided in Table 6.

Table 6: Participant Distribution by Gender and Technical Background in Listening Tests V1 and V2

	Listening Test V1	**Listening Test V2**
Gender		
Male	6	7
Female	18	15
Technical Background		
Yes	8	11
No	16	11
Total	24	22

3.6.2 Listening test

The larger-scale listening tests[4] were conducted using Qualtrics (2005) to evaluate friendliness perception in synthetic speech (refer to Appendix B.1 and Appendix B.2 for detailed listening tests and their translations). The tests were distributed online and participants voluntarily completed them using their preferred device, without supervision. Prior to the listening test,

[4]The listening test Version 1 can be found at https://rug.eu.qualtrics.com/jfe/form/SV_9Kr2dgVGsCs8Zfg; Version 2 can be found at https://rug.eu.qualtrics.com/jfe/form/SV_40zzquXZjksjDVk.

participants were presented with a brief explanation about the research objectives and the procedure. They were then asked to provide consent. The consent statement outlines the voluntary nature of the study, the ability to withdraw participation at any time, permission to use their responses for research purposes, and the data storage procedures in compliance with GDPR regulations. Upon agreeing to voluntary participation, some short questions about demographic information were provided, including native languages, gender, and background in the speech tech field. Participants' native languages were requested because the acoustic manifestation of attitudinal speech, including a friendly attitude, is language-specific in both production and perception as discussed in subsection 2.3. Therefore, perceiving friendliness in Chinese synthetic speech requires native Chinese Mandarin speakers. The gender information was also requested following insights from a study by Kim and Kim (n.d.), which indicated significant differences in speech perception between males and females in certain cases, while not in others. Moreover, there is a lack of research investigating gender differences specifically in attitudinal speech perception, thus warranting further investigation. Furthermore, we also inquired participants' backgrounds in the speech technology-related field. The backgrounds of some participants, such as fellow students specializing in voice technology, might influence the results. Because they may analyze the synthesized voices from a more technical perspective compared to the general population, potentially introducing bias. Therefore, a mixed-group approach where some participants are from the voice tech field while others are not was adopted, allowing for a comparative analysis between participants with a speech tech background and those without.

After answering these questions, three example sentences were provided to participants to familiarize them with the synthetic voices and to adjust the volume to a suitable level for the test. Given that intensity is a key acoustic feature being examined, it is crucial to maintain a consistent volume throughout the entire listening test. Consequently, participants were instructed to set up the volume during this step and were advised not to alter it thereafter. The example sentences chosen are "今天的夜空很美" (The night sky is beautiful tonight), "我明天要去学校上课" (I have to go to school for classes tomorrow), and "我喜欢运动" (I like sports). These sentences were designed to be neutral and devoid of emotional charge, following the same approach used in stimuli design.

The questions and procedures were identical across both listening tests up to this point, with subsequent questions differing in each version. Prior to presenting the main questions, participants were given an example question. This example mirrored the format and options of the main questions in each version, aiming to familiarize participants with the structure of the questions. Following this, the formal test commenced, consisting of 10 sentences randomly chosen from a pool of 30 sentences as detailed in Appendix A. These 10 sentences were displayed in a randomized order for each participant, with each sentence accompanied by three audio clips generated under three conditions: N, P, and PED. The presentation of audio clips under the three conditions was also randomized. Listeners could listen to the audio samples by clicking on buttons and were allowed to listen to them multiple times. Following the audio playback, listeners encountered different types of questions depending on the test version. In version 1, they were required to perform a forced-choice identification task where they selected the perceived attitude from four predefined options: friendly, authoritative, distant, and neutral. This choice of four different tones, including a friendly one, was provided to prevent the priming effect. In version 2, listeners rated how friendly the audio sounded on a scale of 1-5. This rating of perceived friendliness was intentionally designed to prime participants and to provide extra contextual cues. All questions had to be answered before proceeding to the next page, as once they moved on to the next page, they could not return to modify their answers. The

entire session typically lasted 10-15 minutes, and listeners underwent no training session and received no feedback about their performance.

3.6.3 Measurements

Participants' responses were measured and evaluated for both listening tests. In the first listening test, accuracy was utilized as a performance metric. Higher accuracies in the P and PED conditions were indicative of a stronger perception of friendliness, while higher accuracies across all three conditions suggested a better ability to discern attitudes. Accuracy, defined as the proportion of correct responses, was calculated for each participant across the three conditions by dividing the number of correct responses by the total number of stimuli (10) in each condition. A response was considered correct if it matched the intended attitude of the stimuli: neutral for the N condition and friendly for the P and PED conditions. Furthermore, the proportion of inaccurate responses across conditions were computed per participant. Subsequently, the average accuracy and the mean distribution of inaccurate responses across all participants were calculated for each condition.

In the second listening test, the MOS was utilized to assess participants' performance. A higher MOS signified a stronger perception of friendliness, whereas a lower MOS suggested a more neutral attitude perception. The MOS was calculated for each participant first, followed by aggregating the MOS scores across all participants for each condition. This dual methodology offers insights into the role of cues in perceiving friendliness.

3.7 Statistical analysis

Nonparametric statistical tests were employed for analysis due to the non-normal data distribution, as indicated by the Shapiro-Wilk normality test results (e.g., Accuracy_PED: W =0.899, $p < .05$; MOS_N: W=0.909, $p < .05$). Specifically, the Wilcoxon signed-rank test was used to assess any statistically significant differences in accuracy for listening test 1 or MOS for listening test 2 across three distinct conditions. Additionally, the Mann-Whitney U Test was applied to examine potential statistically significant differences in accuracy or MOS within each condition based on gender or technical background. This method was selected due to the lack of paired or matched samples between gender and background, resulting from an imbalance among participants in terms of gender and background.

3.8 Ethical considerations

While this research aims to advance expressive and personalized TTS systems by creating voices adaptable to certain social contexts from a neutral voice, it's important to address potential unseen risks. The training of the TTS system involved audio data; however, this study utilized publicly available data AISHELL 3 instead of manually collecting speech data, eliminating ethical concerns associated with personally collected data. The corpus is licensed under Apache License v.2.0, permitting users to modify, distribute, and sublicense the original open source code as well as the dataset. Therefore, any distribution, modification and use of data can be made freely. To evaluate the quality of synthesized voices, human participants were recruited for pilot studies and listening tests. Regarding the ethics of involving human participants, each participant was provided with a consent statement outlining pertinent details of the research and emphasizing the voluntary nature of participation. This allowed participants to understand the research objectives and be aware of their rights to withdraw from the study at

any time. To ensure the replicability of the research, all code were made available via GitHub. The dataset, experimental design, hyperparameters used in model training and sentence synthesis, hardware utilization details, and results were thoroughly documented in the thesis.

This concludes the methodology section, providing a broad overview of the methods utilized in this research. The subsequent section will delve deeper into the results, offering more detailed insights.

4 Results

Two versions of listening tests were conducted, and the results of each test are presented below (all source data are referenced in Appendix C). In each listening test, we examined perceptual differences among all individuals across three conditions, as well as investigated differences in friendliness perception between two gender groups and between those with and without a background in speech technology.

4.1 Listening test 1

In Listening Test 1, we examined mean accuracy as well as the proportion of incorrect responses among all participants across three conditions. Additionally, we compared differences between males and females, as well as between groups with and without expertise in speech technology.

4.1.1 Differences across conditions

The average accuracy in recognizing attitudes across all participants in the N, P, and PED conditions is presented in Figure 4. Participants showed the highest accuracy in the N condition, followed by the P condition, and then the PED condition, with scores of 0.43, 0.29, and 0.28, respectively. A comparison of mean accuracy between paired conditions using the Wilcoxon signed-rank test indicates significant differences between the N and P conditions ($W = 46$, $p < .05$). Conversely, the differences between the N and PED conditions ($W = 60$, $p > .05$) and between the P and PED conditions ($W = 54.4$, $p > .05$) are not significant.

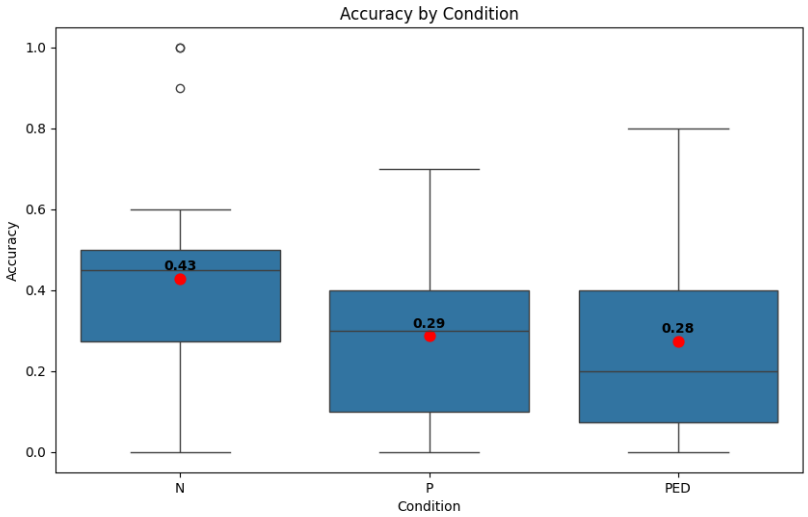


Figure 4: Mean accuracy across all participants in conditions N, P, PED, where red dots represent the mean value.

4.1.2 Distribution of participants' responses

The pattern of inaccurate responses for each condition was also analyzed. Figure 5 presents the distributions of responses—friendly, neutral, distant, and authoritative—across three experimental conditions: N, P, and PED. Each bar chart depicts the mean proportion of these

attitudes within the respective condition. In Condition N, the neutral attitude, which is the correct response, emerges as the most prevalent, accounting for 43% of the responses. The distribution of incorrect responses spans the other three conditions, with distant attitudes representing 25%, friendly attitudes constituting 19%, and authoritative attitudes making up 13% of the responses. Conditions P and PED demonstrate a higher prevalence of identifying friendly attitudes compared to Condition N, constituting 29% and 28% of the responses, respectively. The neutral attitude was most commonly misidentified among incorrect responses, comprising 41% and 33% for conditions P and PED respectively, with the other two attitudes also accounting for a significant portion of responses. In Condition P, friendly attitudes were incorrectly identified as distant with a 13% chance and as authoritative with a 18% chance. In Condition PED, distant and authoritative attitudes each comprised 19% and 20% of responses.

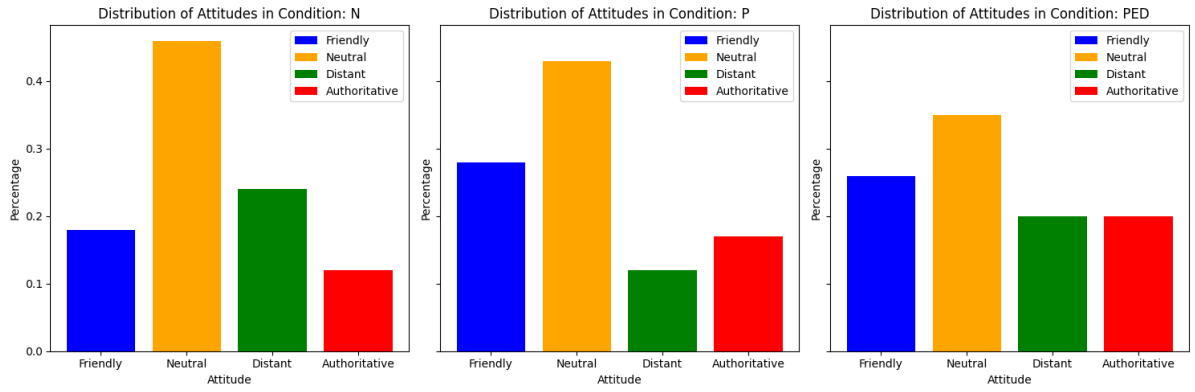


Figure 5: The distributions of responses across three conditions.

4.1.3 Background differences

The impact of participants' expertise in speech technology on accurate attitude identification was evaluated across three conditions (N, P, PED). As shown in Figure 6, there are differences in mean accuracy between participants with (Y) and without (N) a background in speech technology across conditions. In condition N, participants lacking speech technology expertise demonstrated higher accuracy (0.45) than those with a background (0.38). Conversely, participants without expertise exhibited lower mean accuracy in conditions P and PED compared to their counterparts. In the P condition, the mean accuracy for participants with a background in speech technology was 0.39 compared to 0.24 for those without. Similarly, in the PED condition, the mean accuracy for participants with a speech technology background was 0.33, higher than 0.25 for those without. However, statistical analysis revealed these differences to be insignificant across all conditions (N: W = 48.5, p $>$.05; P: W = 90.5, p $>$.05; PED: W = 72.5, p $>$.05). The box plots in Figure 5 also illustrate the variability within each group. It is noteworthy that participants with speech technology expertise display more consistent accuracy compared to those without such expertise, showing less variation across the three conditions.

4.1.4 Gender differences

The impact of gender on accurately recognizing attitudes was investigated across three conditions. Figure 7 illustrates the mean accuracy for female (F) and male (M) participants. In the N condition, female participants demonstrated a slightly lower mean accuracy compared to male participants, with mean accuracy scores of 0.41 for females and 0.48 for males. On the

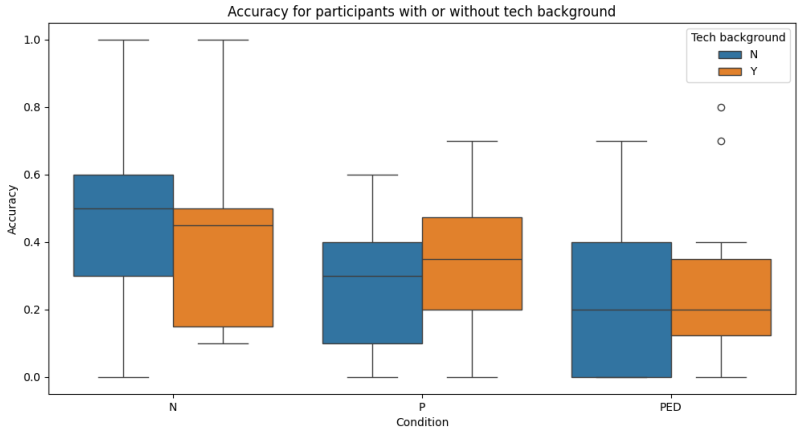


Figure 6: Mean Accuracy of Participants in Conditions N, P, and PED, conditioned by Speech Technology Background.

contrary, female participants exhibited higher mean accuracies than male participants in the P condition and PED condition. For the P condition, mean accuracy was 0.33 for females and 0.17 for males. In the PED condition, females scored 0.29 in accuracy, while males scored 0.22. Furthermore, females showed less variation across the three conditions than males, suggesting more consistent accuracy regardless of the condition. However, these gender differences were statistically insignificant across all three conditions (N: W = 62, p $>$.05; P: W = 35.5, p $>$.05; PED: W = 36.5, p $>$.05).

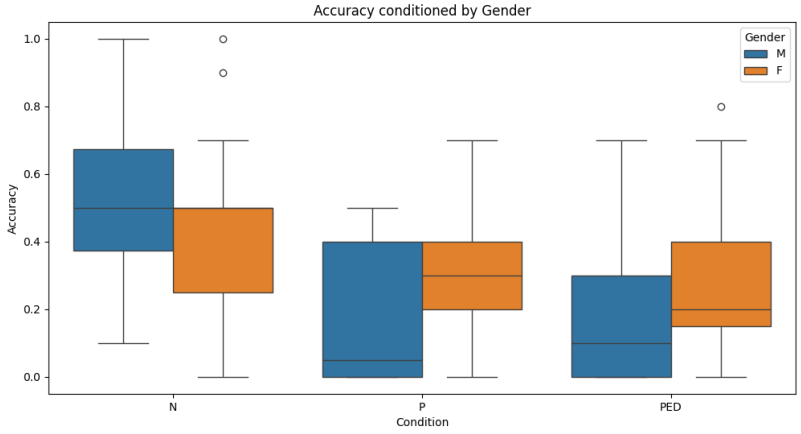


Figure 7: Mean Accuracy of Participants in Conditions N, P, and PED, conditioned by gender.

4.2 Listening test 2

In Listening Test 2, we examined participants' friendliness perception using MOS. Similar to Listening Test 1, we also compared participants' performance across three conditions, between gender groups, and between two expertise groups.

4.2.1 Differences across conditions

The MOS of perceived friendliness across all participants in the N, P and PED conditions was presented in Figure 8. Participants showed the highest MOS in the condition P, followed by the condition PED and then condition N, with scores of 2.74, 2.6 and 2.5 respectively. A comparison of MOS between paired conditions indicates insignificant differences across all pairs: between the N and P conditions (W = 92.5, p >.05), between condition N and PED (W = 97.5, p >.05), and between conditions P and PED (W = 63.5, p >.05).

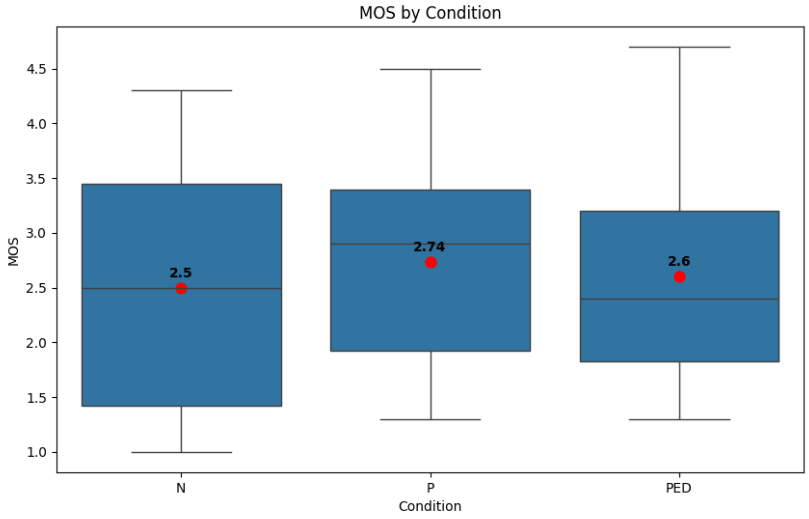


Figure 8: MOS across all participants in conditions N, P, PED, where red dots represent the mean value.

4.2.2 Background Differences

The influence of participants' expertise in speech technology on their perception of friendliness was assessed across three conditions (N, P, PED). Figure 9 displays participants' MOS across three conditions, categorized by whether participants have a background in speech technology (indicated by "Y") or not (indicated by "N"). In condition N, participants without a speech technology background have a MOS of 2.0 while those with this background have a higher MOS of 3.0. This pattern persisted in conditions P and PED, where participants without expertise in speech technology had MOS of 2.5 and 2.4, respectively, whereas those with expertise scored 3.0 and 2.8 in corresponding conditions. However, no statistically significant differences were observed in these conditions (N: $W = 79.5, p > .05$; P: $W = 75.5, p > .05$; PED: $W = 73.5, p > .05$). Additionally, the box plots in Figure 9 depict the variability within each group, highlighting that participants with speech technology expertise exhibited less variation in accuracy across the three conditions.

4.2.3 Gender Differences

The impact of gender on rating perceived friendliness, measured by MOS, was examined across three conditions as depicted in Figure 10. Across all conditions, male participants displayed slightly higher MOS compared to female participants. Specifically, females scored 2.4, 2.7, and 2.5, while males scored 2.8, 2.8, and 2.7 in conditions N, P, and PED respectively. Additionally,

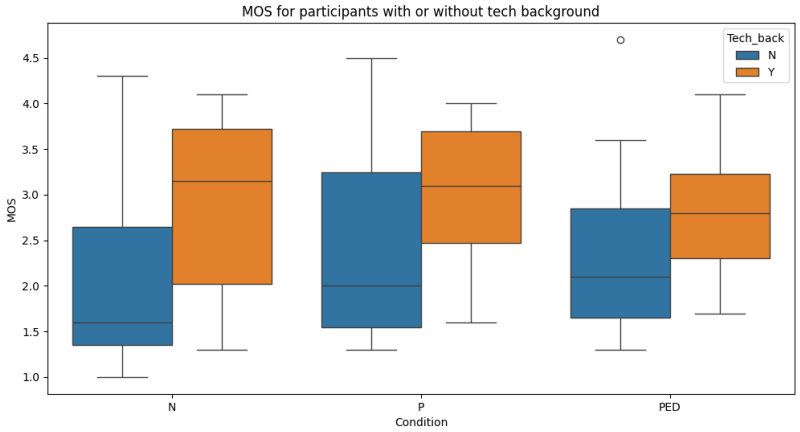


Figure 9: MOS of Participants in Conditions N, P, and PED, conditioned by Speech Technology Background.

it was observed that males showed less variability in ratings across the three conditions compared to females, indicating more consistent ratings. However, statistical analysis revealed that these differences were not significant between the two gender groups across all three conditions (N: $W = 64.5, p > .05$; P: $W = 61, p > .05$; PED: $W = 55.5, p > .05$).

This section presents the results obtained from two versions of listening tests, accompanied by plots illustrating variations in participants' perception of friendliness across distinct conditions, gender, and background. The subsequent section will delve into a comprehensive discussion of these results, elucidating how they address the research question and validate the hypotheses.

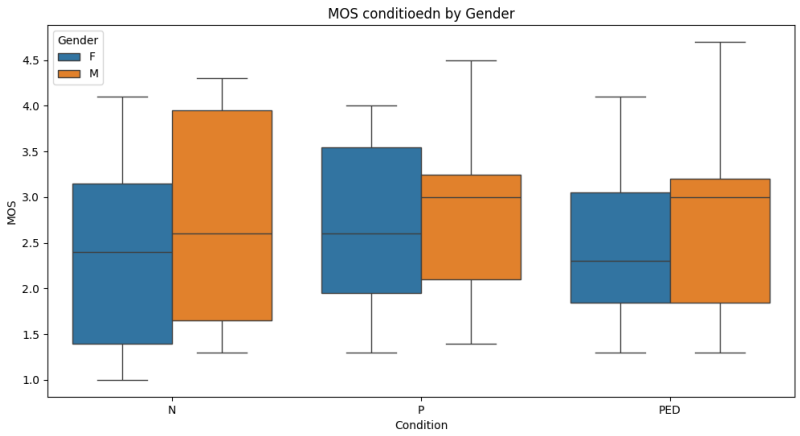


Figure 10: MOS of Participants in Conditions N, P, and PED, conditioned by gender.

5 Discussion

The main objective of this study is to synthesize friendly speech in Mandarin Chinese by manipulating acoustic attributes in neutral speech using the FastSpeech 2 (FS2) framework, thereby achieving a perceptual shift from a neutral to a friendly vocal persona. This section addresses the main research question: how do specific manipulations of pitch, duration, and energy impact the perception of friendliness in synthetic stimuli? Additionally, it validates two hypotheses. First, the study explores whether increasing the pitch of neutral speech alone significantly enhances the perception of friendliness. Second, it investigates whether modulating a combination of acoustic features, including pitch, duration, and energy, leads to an optimal perception of friendliness.

5.1 Validation of the First Hypothesis

When comparing friendliness perception between conditions N and P, the first hypothesis that increasing the mean pitch of a neutral voice significantly enhances the perception of friendliness is called into question. In listening test 1, although the perception of friendliness increased from 18% in condition N to 29% in condition P, participants accurately identified a friendly attitude only 29% of the time, with listeners perceiving the intended friendly attitude as neutral in over 40% of cases under condition P. A low accuracy rate and a neutral attitude misidentification ratio above 40% suggest that pitch modulation does not significantly alter the perception of a neutral attitude to a friendly one, thereby causing participants' difficulty in distinguishing between neutral and friendly attitudes under condition P. In addition to the more frequent perception of neutrality under condition P, the distribution of inaccurate responses across the other two attitudes, with 12% for distant attitudes and 17% for authoritative attitudes, suggests participants' challenge in distinguishing among different social attitudes in general. As shown in subsection 4.1.2, listeners incorrectly identify friendly attitudes as a distant attitude with a 12% chance and as an duthoritative attitude with a 17% chance. Referring to subsection 2.3, authoritative speech is associated with a lower mean pitch compared with neutral speech cross-linguistically, including in Mandarin Chinese. Similarly, a distant attitude is also conveyed by low pitch according to Salais et al. (2022). Therefore, it is surprising to observe that participants recognized the intended friendly attitude with increased pitch in condition P as distant and authoritative speech characterized by low pitch. The difficulty in distinguishing between acoustically contrasting social attitudes in pitch may suggest that participants have a general difficulty in recognizing nuanced social attitudes, rather than being due to similarities in the attitude-specific acoustic correlates.

In listening test 2, the MOS of perceived friendliness in condition P slightly exceeds that in condition N, indicating a slight enhancement in the perception of a friendly attitude following pitch modulation. This more pronounced impact of upward pitch manipulation in listening test 2 utilized the same synthetic stimuli as in listening test 1; the only distinction between the two tests lies in their experimental design, with listening test 2 intentionally incorporating a priming effect to provide participants with additional contextual information beyond acoustic cues. Consequently, the findings may suggest that extra contextual cues enhance the perception of friendliness and cues from multiple sources facilitate listeners in discerning nuances in social attitudes. Nevertheless, the effect of additional cues in contexts might be limited as pitch modulation fails to yield a significantly heightened perception of friendliness in listening test 2.

Overall, the upward manipulation of pitch does not significantly enhance participants' per-

ception of friendliness across two listening tests, thereby challenging the first hypothesis.

5.2 Validation of the Second Hypothesis

The second hypothesis, which posited that the combined effects of acoustic manipulation on pitch, duration, and energy in neutral speech would further enhance friendliness perception compared to adjusting pitch alone, was not confirmed. Findings from both listening tests challenged this hypothesis. In Listening Test 1, the accuracy in identifying friendly attitudes dropped from 29% in condition P to 28% in condition PED. Likewise, in Listening Test 2, the MOS for perceived friendliness decreased from 2.74 in condition P to 2.60 in condition PED. This decline in perceived friendliness in the PED condition suggests that additional acoustic modulations in duration and energy degraded friendliness perception rather than enhancing it. Given that energy modulation was minimal, the degradation is likely caused by the shortened duration in the PED condition. The stimuli used in this study were short, ranging from 0.75 seconds to 1.67 seconds in condition N, and further shortening in the PED condition may have led to utterances that were too short. When samples are excessively short, participants may find it more challenging to discern social attitudes, thereby increasing reliance on pure chance during the identification process. This is evident from the apparent confusion participants experienced regarding the different social attitudes, particularly in the PED condition, where responses were similarly distributed across the four attitudes. However, this explanation requires further investigation. Moreover, the distribution of incorrect responses in Listening Test 1 provides additional insights into potential reasons for the decreased perception of friendliness. Participants more frequently misrecognized friendly attitudes as distant attitudes in the PED condition than in the P condition. This pattern of misidentification can be attributed to the fact that a distant attitude is characterized by fast speech (Salais et al., 2022), thus, the shortened duration in the PED condition may have led to increased confusion between a friendly attitude and a distant attitude. Additionally, participants still inaccurately identified the friendly attitude as neutral one-third of the time, suggesting that manipulating pitch, duration, and energy together still fails to achieve a more friendly transition from neutral speech.

Neither modulating individual acoustic feature pitch nor manipulating combined acoustic correlates of pitch, duration, and energy in neutral speech led to a significantly friendlier perception. These results deviate from previous studies on synthesizing friendly speech through acoustic modulations in neutral speech, which reported significantly higher friendliness perception following pitch manipulation and optimal friendliness perceived through the combined manipulation of pitch, duration, and energy (Li et al., 2004; Li & Wang, 2004), as discussed in subsection 2.5. Discrepancies in the results may stem from differences in the dataset used or the acoustic control values employed. While the previous two studies utilized natural speech data, this study utilized read speech data from the dataset AISHELL 3 and performed acoustic modulations based on observed acoustic differences between friendly and neutral speech produced by two drama school students in the reference study by F. Chen et al. (2004). However, acoustic differences observed in the reference paper were based on speech produced by two male speakers from a drama school. It is noteworthy that vocalizations from voice actors are often less authentic than spontaneous speech (Anikin & Lima, 2018; Salais et al., 2022), and genuine experiences can induce more substantial physiological changes which could affect vocal characteristics in speech production (Niedenthal, Winkielman, Mondillon, & Vermeulen, 2009). Therefore, reliance on data performed by voice actors introduces potential confounding factors, as the acoustic differences observed in the reference study may not entirely mirror those found in natural speech.

5.3 Cue impact

Differing from previous studies that successfully transitioned neutral speech to friendly speech after acoustic manipulating, this study revealed that participants struggled not only with perceiving friendliness from synthetic stimuli in both conditions P and PED, but also with distinguishing among social attitudes. This difficulty may result from the insufficiency of cues to help listeners to perceive friendliness from speech. In Listening Test 1, participants experienced significant trouble identifying friendly attitudes based solely on acoustic cues. Although their ability to perceive friendliness improved when additional contextual cues were included in Listening Test 2, the enhancement was limited. Mehrabian and Wiener (1967) and Zhou, Sisman, Liu, and Li (2022) suggest that linguistic messages and nonverbal paralinguistic features account for only 7% and 38% of the communication of social attitudes, respectively, while facial cues contribute 55%, highlighting the role of visual cues in perceiving social attitudes. Barbulescu, Ronfard, and Bailly (2017) highlighted the importance of visual information in perceiving expressions of attitudes from the perspective of speech production. They found that male speakers express attitudes utilizing energy, head movements, gaze and upper-face expressions, while female speakers rely on F0 and lower-face expressions. Since acoustic manipulation was performed on speech produced by male speakers in this study, the absence of visual information prevents a significant amount of friendly cues expressed through speakers' head movements, gaze, and upper-face expressions, thereby impeding participants' accurate perception of attitudes. Furthermore, the McGurk effect (McGurk & MacDonald, 1976) provides compelling evidence that visual information leads a dominant position in some speech perception. The McGurk effect is an audiovisual illusion where non-matching visual cues greatly alter the perception of sounds when they conflict with auditory cues (Kim & Kim, n.d.; McGurk & MacDonald, 1976; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007). For example, listeners tend to perceive a "Fa" sound when the speaker's lip movement shows "Fa," even if the original sound is "Ba" (Kim & Kim, n.d.). We often use different sensory information from different modalities (e.g., audition and vision) to perceive speech in everyday conversations (Kim & Kim, n.d.), the absence of visual information in this study may have hindered participants' ability to accurately perceive attitudes.

5.4 Expertise effects

Considering the minimal variations in friendliness perception among all participants, we proceeded to investigate the differences within specific groups, starting with those with and without expertise. In Listening Test 1, the results suggest that participants with a background in speech technology may have an advantage in recognizing friendly attitudes, as evidenced by their higher accuracies in the P and PED conditions compared to those without expertise. Similarly, in Listening Test 2, participants with speech technology expertise demonstrated higher MOS than their counterparts across all three conditions, indicating a stronger perception of friendliness. These findings suggest that familiarity with speech technology can improve the ability to perceive friendliness in synthetic speech and enhance the recognition of nuanced social attitudes. This aligns with previous research on the impact of musical expertise on emotion perception in speech.

Music and speech share many acoustic characteristics, including spectral envelope, duration, and fundamental frequency. In music, the spectral envelope influences a sound's timbre, while pitch and rhythm shape the melody. In speech, the spectral envelope aids in distinguishing vowels, and pitch and rhythm shape prosody. Additionally, our perceptual system excels at

isolating individual auditory objects from background noise, such as discerning a conversation partner's voice in a crowd or picking out an instrument in an orchestra. These commonalities suggest a similar auditory processing mechanism between the two domains (Varnet, Wang, Peter, Meunier, & Hoen, 2015). Therefore, studies on musical expertise shed light on the influence of speech field experience on attitude perception in speech. Thompson, Schellenberg, and Husain (2004) found that musically trained adults outperformed untrained adults in identifying four basic emotions in speech, suggesting that music experience enhances sensitivity to emotions. Similarly, Lima and Castro (2011) reported a robust effect of expertise that musicians were more accurate in recognizing emotions in speech prosody compared to non-musicians. The effect of musical experience on emotion perception may be attributed to its impact on pitch processing in speech. This is supported by electrophysiological evidence, which suggests that music training affects pitch processing in speech, and pitch is a key acoustic feature for distinguishing emotions in both music and speech (Lima & Castro, 2011). Similarly, expertise in speech technology may affect pitch processing and thereby influence the recognition of attitudes.

However, no statistically significant differences were observed in either accuracy in listening test 1 or MOS in listening test 2 between participants with and without expertise. According to Lima and Castro (2011), both musicians and non-musicians perceived the general emotional properties of the stimuli similarly, with no discrepancies in misclassification patterns or in the acoustic cues influencing their categorization responses. This similarity suggests that musicians' higher accuracy in recognizing emotions probably reflects quantitative rather than qualitative differences in processing emotional prosody. Furthermore, Lima and Castro (2011) argued that positive effects of musical training on emotion perception are closely linked to the level of musical training, with longer years of expertise showing more profound effects, as extensive training is necessary to detect experience-related behavioral differences such as the recognition of emotions through voice. Therefore, the limited effect of expertise on attitude perception in this study might be attributed to the fact that participants with experience in the speech field were recruited from a master's program where all students were exposed to speech technology for only a year. Thus, limited training in the field may not be sufficient to lead to significant qualitative differences in identifying nuanced social attitudes.

5.5 Gender differences

The differences in perceiving friendly attitudes between gender groups was also examined. In Listening Test 1, female participants demonstrated higher accuracies in recognizing friendly attitudes in conditions P and PED compared to male participants, suggesting a superior capability in identifying speech attitudes. This finding aligns with the conclusion proposed by Kim and Kim (n.d.), who conducted a systematic analysis on how gender differences affect speech perception, suggesting that females exhibit better abilities and greater sensitivity to speech stimuli. Unexpectedly, male participants showed higher MOS on friendliness perception than female participants in listening test 2. This unexpected outcome may be attributed to the limited number of male participants in Listening Test 2. Only seven males participated in the listening test 2 and two of them scored significantly higher than other male participants across all conditions. The limited number of male participants makes the data sensitive to extreme values. Therefore, further investigation with a larger sample size and a more balanced gender distribution is necessary to validate these findings.

While there are observable differences in both accuracy and MOS distributions between genders, these differences were not statistically significant. This suggests that gender does

not decisively influence participants' perception of attitudes under the tested conditions. This finding aligns with Ponsot et al. (2018), who reported that male and female listeners judged dominance and trustworthiness in short utterances similarly. According to Kim and Kim (n.d.), previous research has found both gender differences and similarities in speech perception, with the presence or absence of gender effects potentially linked to task types, stimuli used, and the biological cycles of perceivers. Therefore, further investigation into the impact of gender on the perception of friendly attitudes, with controlled task types, stimuli, and biological factors, is necessary.

5.6 Limitations

The study's results offered insights into the research questions, although they did not align with the anticipated outcomes. However, it is important to acknowledge several limitations. Firstly, the study's participant pool was restricted in both size and demographic composition, primarily consisting of educated individuals aged 20 to 35 recruited from personal networks. Therefore, the findings may not be fully representative of the broader population, thereby limiting the generalizability of the results. Secondly, because of the scarcity of expressive speech data, particularly regarding friendly attitudes in Mandarin Chinese, and time constraints, we were unable to collect natural friendly speech data from either a dataset or manual sources. Consequently, we could not include natural speech in the listening test to serve as a reference. Therefore, participants' inability to perceive friendliness from synthesized friendly speech may stem from their general difficulty in discerning nuanced attitudinal speech, even in natural speech, or from the unsuccessful acoustic manipulations of neutral speech data. It is challenging to determine the exact reasons. These limitations underscore the necessity for cautious interpretation of the findings and emphasize the importance of addressing methodological gaps in future research.

In summary, while my initial hypotheses have not been confirmed, the research question has been addressed and the research objectives have been achieved. The subsequent section will offer a conclusion, summarizing the key findings and their implications, and suggesting future directions for research in this specific topic.

6 Conclusion

This section will offer a concise overview of the main findings of the research, followed by an outline of future directions and potential avenues. It will conclude with a subsection discussing both the academic significance and industrial relevance of this research.

6.1 Summary

This research synthesizes speech with a friendly attitude by manipulating acoustic features in neutral speech to achieve a perceptual transition from neutral to friendly personas. However, altering the pitch alone did not result in a significant shift to friendly speech. Furthermore, incorporating shortened duration and slightly increased energy also failed to enhance the perception of friendliness. The insignificant differences between the original neutral speech and the acoustically modulated speech can be attributed to several factors. Firstly, the dataset comprised read speech and the acoustic manipulation values were derived from findings obtained from acted speech, which may be less authentic than spontaneous speech data. This lack of authenticity could lead to less pronounced changes in speech production, subsequently affecting perception. Secondly, visual information is a crucial cue for perceiving social attitudes, and the absence of visual cues in this study may have hindered participants' ability to perceive friendliness.

6.2 Future Work

The conclusions drawn in this study are based on acoustic analysis of acted speech, which differs from spontaneous speech. Therefore, it is crucial to validate the findings of this study using spontaneous speech data. This can be accomplished by conducting acoustic analyses on both genuine friendly and neutral speech in the future work. Subsequently, manual manipulation of acoustic features can be conducted to transform neutral speech into friendly speech, according to the acoustic differences found in the analysis. This approach provides empirical evidence to either support or challenge the conclusions drawn from the study.

This study modulated acoustic features to transition from neutral speech to friendly speech; however, this indirect method did not achieve a successful perception transition. Future research could synthesize friendly speech directly using an expressive TTS system with style tokens (such as Tacotron2) and employing expressive friendly speech data, followed by a perception test to evaluate the perception of friendliness. By comparing perception results from directly and indirectly synthesized friendly speech, we can examine the effectiveness of the direct approach relative to the indirect approach. This comparison will help determine whether the failure to achieve successful friendliness perception in this study is due to the indirect approach or if participants are generally insensitive to nuanced social attitudes based solely on acoustic cues.

Additionally, it is worth conducting a multi-modal study on friendliness perception. Previous research suggests that facial cues account for 55% of the communication of social attitudes, while nonverbal paralinguistic features contribute 38% (Mehrabian & Wiener, 1967; Zhou et al., 2022). A comparative study between sole acoustic cues and a combination of acoustic and visual cues can investigate whether acoustics alone are insufficient for nuanced friendly attitude perception and if the inclusion of visual cues significantly enhances the perception of friendliness.

Furthermore, this study exclusively examines perceptions of friendliness using manipulated neutral speech data. Future research can expand the scope of investigation to include a broader range of vocal personas beyond friendly one, using the same methodology. This extension could establish mechanisms for transitioning between vocal personas via acoustic adjustments and explore connections between various vocal personas, complementing the framework proposed by Noufi et al. (2023).

Moreover, this extension could involve exploring different vocal personas across diverse linguistic and cultural contexts. Noufi et al. (2023) argued that the adoption of a vocal persona is shaped by an individual's orientation toward a specific language or culture. They propose that multilingual speakers modify acoustic features, such as pitch range and vocal tone, beyond what is linguistically necessary when conversing in a non-native language, leading their conversational partners to perceive a completely different persona. Future research can investigate the interaction between speakers' second languages and vocal personas, revealing how individuals adjust their vocal expression when communicating in a second language.

By conducting future research on these areas, a more comprehensive understanding of vocal persona and a more expressive TTS system involving social attitudes can be achieved.

6.3 Impact and relevance

Noufi et al. (2023) highlighted the fluidity of vocal personas and the significance of contextualized environments in vocal persona transition. However, their study did not delve into the mechanism for transitioning from one vocal persona to another via acoustic adjustments. This study examines precise control over acoustic features to transition from a neutral to a friendly vocal persona, with the potential to expand this transition to other vocal personas beyond just friendly ones in the future. As a result, this study not only extends the vocal persona framework proposed by Noufi et al. (2023) but also offers more nuanced contexts for comprehending vocal persona transitions.

While most studies have examined between-individual differences in nonverbal communication, few researchers have investigated the within-individual variations in voice (Sorokowski et al., 2019), especially in terms of how individuals adjust their vocal personas and modulate key vocal parameters across different social settings. By exploring within-individual voice transition to align with social contexts, this study holds theoretical significance by advancing our understanding of how individuals dynamically navigate social environments through vocal adaptation, contributing to the broader discourse on human social behavior and communicative processes.

Beyond academic significance, this study also aims to contribute to industrial implications. It employs acoustical manipulations in pitch, duration, and energy to transform neutral speech into friendly speech, rather than relying on expressive friendly speech for synthesis. This approach is particularly valuable due to the limited availability of expressive speech data, especially in attitudinal speech, and the substantial costs associated with data collection, annotation, and validation. With the potential to synthesize expressive speech from existing neutral speech data, it may effectively tackle the challenges of data scarcity and high costs.

This study synthesizes expressive friendly speech from neutral speech, catering to casual interactions with friends and family, as well as specific professional settings with colleagues and clients. It has the potential to synthesize other forms of expressive speech across diverse social contexts in the future, such as authoritative speech for professional environments and dominant speech to convey hierarchical social status.These synthetic voices can represent a variety of vocal identities and social presences across different communication scenarios, potentially

addressing the challenge of expressivity and advancing personalized synthesized speech. This advancement can benefit individuals with voice impairments, companies seeking voice branding strategies, and industries involved in dubbing video games, films, and audiobooks, allowing them to present desired characters or social presence effectively to the audiences. Furthermore, the study offers benefits to virtual assistants, human-computer interaction (HCI), and voice-user interfaces (VUIs), enabling them to communicate with users expressively and convey messages more effectively.

References

Anikin, A., & Lima, C. F. (2018). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Quarterly Journal of Experimental Psychology*, *71*(3), 622–641.

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

Aucouturier, J.-J., & Canonne, C. (2017). Musical friends and foes: The social cognition of affiliation and control in improvised interactions. *Cognition*, *161*, 94–108.

Aung, T., & Puts, D. (2020). Voice pitch: a window into the communication of social power. *Current opinion in psychology*, *33*, 154–161.

Barbulescu, A., Ronfard, R., & Bailly, G. (2017). Which prosodic features contribute to the recognition of dramatic attitudes? *Speech Communication*, *95*, 78–86.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current directions in psychological science*, *20*(5), 286–290.

Callier, P. R. (2013). *Linguistic context and the social meaning of voice quality variation*. Georgetown University.

Chen, A., Rietveld, T., & Gussenhoven, C. (2001). Language-specific effects of pitch range on the perception of universal intonational meaning. In *7th european conference on speech communication and technology (eurospeech 2001)* (pp. 1403–1406).

Chen, F., Li, A., Wang, H., Wang, T., & Fang, Q. (2004). Acoustic analysis of friendly speech. In *2004 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. I–569).

Geng, P., Gu, W., Johnson, K., & Erickson, D. (2020). Acoustic-prosodic and articulatory characteristics of the mandarin speech conveying dominance or submissiveness. In *Proc. 10th international conference on speech prosody* (pp. 424–428).

Henter, G. E., Lorenzo-Trueba, J., Wang, X., & Yamagishi, J. (2017). Principles for learning controllable tts from annotated and latent variation. In *Interspeech* (pp. 3956–3960).

Hildebrand-Edgar, N. (2016). *Creaky voice: An interactional resource for indexing authority* (Unpublished doctoral dissertation).

House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in swedish human–machine dialogue. *Speech Communication*, *46*(3-4), 268–283.

Huang, K.-L., Duan, S.-F., & Lyu, X. (2021). Affective voice interaction and artificial intelligence: A research study on the acoustic features of gender and the emotional states of the pad model. *Frontiers in Psychology*, *12*, 664925.

Ivanov, A. V., Riccardi, G., Sporka, A. J., & Franc, J. (2011). Recognition of personality traits from human spoken conversations. In *Twelfth annual conference of the international speech communication association.*

Kim, A., & Kim, L. (n.d.). A systematic analysis of speech perception and sex differences.

Lee, K., Park, K., & Kim, D. (2021). Styler: Style factor modeling with rapidity and robustness via speech decomposition for expressive and controllable neural text to speech. *arXiv preprint arXiv:2103.09474*.

Li, A., Chen, F., Wang, H., & Wang, T. (2004). Perception on synthesized friendly standard chinese speech. In *International symposium on tonal aspects of languages: With emphasis on tone languages.*

Li, A., & Wang, H. (2004). Friendly speech analysis and perception in standard chinese. In *Eighth international conference on spoken language processing.*

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*(5), 1021.

Liu, W., Zhang, X., & Liang, C. (2023). An acoustic study on character voices of dominators and subordinates: A case study on male characters in empresses in the palace. *Frontiers in Communication*, *7*, 1088170.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? personality impressions from brief novel voices. *PloS one*, *9*(3), e90779.

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

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of personality and social psychology*, *6*(1), 109.

Meier, P. (n.d.). Creaky voice in american english: How are american women who use creaky voice perceived? a literature review. *Leviathan: Interdisciplinary Journal in English*(9).

Moine, C. L., & Obin, N. (2020). Att-hack: An expressive speech database with social attitudes. *arXiv preprint arXiv:2004.04410*.

Niedenthal, P. M., Winkielman, P., Mondillon, L., & Vermeulen, N. (2009). Embodiment of emotion concepts. *Journal of personality and social psychology*, *96*(6), 1120.

Noble, L., & Xu, Y. (2011). Friendly speech and happy speech-are they the same? In *Icphs* (pp. 1502–1505).

Noufi, C., May, L., & Berger, J. (2023). Context, perception, production: A model of vocal persona. *PsyArXiv. July*, *28*.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of f of voice. *Phonetica*, *41*(1), 1–16.

Omar, M. K., & Pelecanos, J. (2010). A novel approach to detecting non-native speakers and their native language. In *2010 ieee international conference on acoustics, speech and signal processing* (pp. 4398–4401).

Park, K., & Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: a window into the origins of human vocal control? *Trends in cognitive sciences*, *20*(4), 304–318.

Polzehl, T., Moller, S., & Metze, F. (2011). Modeling speaker personality using voice.

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, *115*(15), 3972–3977.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior*, *27*(4), 283–296.

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, *27*(1), 89–115.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

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

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the mcgurk illusion. *Neuropsychologia*, *45*(3), 587–597.

Salais, L., Arias, P., Le Moine, C., Rosi, V., Teytaut, Y., Obin, N., & Roebel, A. (2022). Production strategies of vocal attitudes. In *Interspeech 2022* (pp. 4985–4989).

Scherer, K. R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index. *Journal of Nonverbal Behavior*, *35*, 305–326.

Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in cognitive sciences*, *21*(3), 216–228.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2010). The interspeech 2010 paralinguistic challenge. In *Proc. interspeech 2010, makuhari, japan* (pp. 2794–2797).

Shi, Y., Bu, H., Xu, X., Zhang, S., & Li, M. (2020). Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Sicoli, M. A. (2010). Shifting voices with participant roles: Voice qualities and speech registers in mesoamerica. *Language in Society*, *39*(4), 521–553.

Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., ... Pisanski, K. (2019). Voice of authority: Professionals lower their vocal frequencies when giving expert advice. *Journal of Nonverbal Behavior*, *43*, 257–269.

Tagg, P. (2012). *Music's meanings: a modern musicology for non-musos*. Mass Media's Scholar's Press.

Tang, P., & Gu, W. (2015). Perceptual experiment and acoustic analysis of chinese attitudes: A preliminary study. In *Icphs*.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, *4*(1), 46.

Tits, N., El Haddad, K., & Dutoit, T. (2021). Analysis and assessment of controllability of an expressive deep learning-based tts system. In *Informatics* (Vol. 8, p. 84).

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015). How musical expertise shapes speech perception: evidence from auditory classification images. *Scientific reports*, *5*(1), 14489.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., ... Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning* (pp. 5180–5189).

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile american women? *American Speech*, *85*(3), 315–337.

Zäske, R., Schweinberger, S. R., & Kawahara, H. (2010). Voice aftereffects of adaptation to speaker identity. *Hearing Research*, *268*(1-2), 38–45.

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, *137*, 1–18.

Appendices

A Sentences for synthesis

Index	Sentence	English translation	P condition	PED condition		
			pitch control	pitch control	duration control	energy control
1	这是真的	it's true	1.2475	1.2475	0.9403	1.00017
2	现在是中午	it's noon	1.2475	1.2475	0.9403	1.00017
3	你姐姐喝酒了	your sister drank	1.2475	1.3	0.9203	1.00017
4	我明天就要离开	I'm leaving tomorrow	1.2475	1.2475	0.9203	1.00017
5	我要休息一下	I'm taking a break	1.2475	1.2475	0.9203	1.00017
6	给他们一个机会	give them a chance	1.2475	1.2475	0.9203	1.00017
7	我提前离开了	I left early	1.27	1.3	0.9403	1.00017
8	我们唱歌	we sang	1.2475	1.2475	0.9203	1.00017
9	给我一个答案	give me an answer	1.3	1.3	0.9403	1.00017
10	伦敦已经很晚了	it's late in London	1.2475	1.27	0.9203	1.00017
11	纽约还早	it's early in New York	1.2475	1.2475	0.9203	1.00017
12	你也说了同样的话	you said the same thing	1.2475	1.2475	0.9203	1.00017
13	我们要去喝一杯	we're going to have a drink	1.2475	1.2475	0.9203	1.00017
14	你忘了我的夹克	you forgot my jacket	1.2475	1.2475	0.9203	1.00017
15	我离这里很远	I went far from here	1.2475	1.2475	0.9203	1.00017
16	你走得很快	you left quickly	1.2475	1.2475	0.9203	1.00017
17	你去了海滩	you went to the beach	1.2475	1.2475	0.9203	1.00017
18	她去度假了	she went on vacation	1.2475	1.2475	0.9203	1.00017
19	我们稍等一下	we will wait a little	1.2475	1.2475	0.9203	1.00017
20	今天很冷	it was cold today	1.2475	1.2475	0.9203	1.00017
21	伦敦今晚天气很闷	it's heavy tonight in London	1.2475	1.2475	0.9403	1.00017
22	我们要去度假	we are going to take a vacation	1.2475	1.2475	0.9203	1.00017
23	她去了山里	she went to the mountains	1.2475	1.2475	0.9203	1.00017
24	巴黎今晚天气很好	the weather is nice this evening in Paris	1.27	1.2475	0.9203	1.00017
25	她做了我想要做的	she did what I wanted	1.2475	1.2475	0.9203	1.00017
26	你睡了一整晚	you slept all night	1.2475	1.2475	0.9203	1.00017
27	让我们相信他们	let's trust them	1.2475	1.2475	0.9203	1.00017
28	今天我们放松一下	let's take our day	1.2475	1.2475	0.9203	1.00017
29	我们去喝咖啡	let's go have a coffee	1.2475	1.2475	0.9203	1.00017
30	她离开上海了	she left for bale	1.2475	1.2475	0.9403	1.00017

B Listening tests

B.1 Listening test 1

Introduction:

欢迎您参与此次听力实验！通过参与本次实验，您将有机会为研究和改进更具表现力的语音合成技术做出贡献，让合成的声音富有情感，风格和韵律，从而提升用户在语音助手、游戏、有声小说等方面的体验。

Welcome to participate in this listening experiment! By participating in this experiment, you will have the opportunity to contribute to the research and improvement of more expressive speech synthesis technology, allowing synthesized voices to be rich in emotion, style, and rhythm, thereby enhancing users' experiences in areas such as voice assistants, games, and audiobooks.

本次实验预计耗时约为15 分钟，共包含10 个不同的句子。每个句子均配有3 个音频供您聆听。在聆听完音频后，您将需要选择每个音频所表现出的语气风格（不要关注句子意思，而是声音中所表达的语气）。请注意，所有题目没有正确或错误的答案，只需按照您的直觉作答。

This experiment is expected to take about 15 minutes and includes a total of 10 different sentences. Each sentence is accompanied by 3 audio clips for you to listen to. After listening to the audio, you will need to select the attitude expressed by each audio clip (do not focus on the meaning of the sentences, but rather the tone conveyed in the sound). Please note that there are no correct or incorrect answers for all questions, just answer according to your intuition.

本次实验是匿名的，您所提供的答案将不会泄露您的身份信息。如果您同意参与本次试验并允许我们在研究中使用您的结果，请在同意选项中打勾，并在安静的环境中使用耳机开始实验。

This experiment is anonymous, and the answers you provide will not reveal your identity. If you agree to participate in this experiment and allow us to use your results in our research, please check the consent option (attached consent form) and start the experiment in a quiet environment using headphones.

如果您在实验过程中有任何疑问或需要帮助，请随时发送电子邮件至c.lin.22@student.rug.nl。更多细节，请查看此同意书。再次感谢您的支持与参与！

If you have any questions or need assistance during the experiment, please feel free to email c.lin.22@student.rug.nl at any time. For more details, please refer to this consent form. Thank you again for your support and participation!

您同意并愿意继续参与听力实验吗?

Do you agree and are you willing to continue the listening experiment?

❑ 是的，我同意。(Yes, I agree.)　　❑ 不，我不同意。(No, I do not agree.)

Demographic questions

您的母语是中文吗? (Is Chinese your native language?)

❑ 是的(Yes)　　❑ 不是(No)

您的性别是?(What is your gender?)

❑ 男(Male)　　❑ 女(female)

您是否具有与语音技术相关的背景？例如，您是否学习过语音识别或语音合成相关专业，或者在这个领域有工作经验，或者对这方面有深入的了解？(Do you have a

background in speech technology? For example, have you studied majors related to speech recognition or speech synthesis, worked in this field, or have in-depth knowledge in this area?)

❑ 是的(Yes) ❑ 不是(No)

Audio instructions:

这里有3个合成的音频文件，您可以利用它们来调整音量至适中水平，并熟悉生成音频的语音特点。在之后的实验中，请不要调整手机或电脑的音量和音频的播放速度！

Here are three synthesized audio clips that you can use to adjust the volume to a moderate level and familiarize yourself with the characteristics of the synthetic speech. Please do not adjust the volume or playback speed of your phone or computer during the experiment!

Test question:

下面是一个测试，旨在让您熟悉本测试的题型。您的答案不会被记录。

The following is a test question designed to familiarize you with the types of questions in this test. Your answers will not be recorded.

想象一下，如果有人用这样的声音与您说话，您会认为他是用何种语气或态度在与您交谈?

Imagine if someone were to speak to you with such a voice, what tone or attitude would you perceive them to be speaking with?

	友好（Friendly）	权威(Authoritative)	疏离(Distant)	中立(Neutral)
音频1 (Audio1)	❑	❑	❑	❑
音频2 (Audio2)	❑	❑	❑	❑
音频3 (Audio3)	❑	❑	❑	❑

Main question: 10x

接下来是正式的测评内容，请认真回答下列问题。

The following is the formal evaluation content. Please answer the following questions carefully.

想象一下，如果有人用这样的声音与您说话，您会认为他是用何种语气或态度在与您交谈?

Imagine if someone were to speak to you with such a voice, what tone or attitude would you perceive them to be speaking with?

	友好（Friendly）	权威(Authoritative)	疏离(Distant)	中立(Neutral)
音频1 (Audio1)	❑	❑	❑	❑
音频2 (Audio2)	❑	❑	❑	❑
音频3 (Audio3)	❑	❑	❑	❑

B.2 Listening test 2

Introduction:

欢迎您参与此次听力实验！通过参与本次实验，您将有机会为研究和改进更具表现力的语音合成技术做出贡献，让合成的声音富有情感，风格和韵律，从而提升用户在语音助手、游戏、有声小说等方面的体验。

Welcome to participate in this listening experiment! By participating in this experiment, you will have the opportunity to contribute to the research and improvement of more expressive speech synthesis technology, allowing synthesized voices to be rich in emotion, style, and rhythm, thereby enhancing users' experiences in areas such as voice assistants, games, and audiobooks.

本次实验预计耗时约为15 分钟，共包含10 个不同的句子。每个句子均配有3 个音频供您聆听。在聆听完音频后，您将需要选择每个音频所表现出的语气风格（不要关注句子意思，而是声音中所表达的语气）。请注意，所有题目没有正确或错误的答案，只需按照您的直觉作答。

This experiment is expected to take about 15 minutes and includes a total of 10 different sentences. Each sentence is accompanied by 3 audio clips for you to listen to. After listening to the audio, you will need to select the attitude expressed by each audio clip (do not focus on the meaning of the sentences, but rather the tone conveyed in the sound). Please note that there are no correct or incorrect answers for all questions, just answer according to your intuition.

本次实验是匿名的，您所提供的答案将不会泄露您的身份信息。如果您同意参与本次试验并允许我们在研究中使用您的结果，请在同意选项中打勾，并在安静的环境中使用耳机开始实验。

This experiment is anonymous, and the answers you provide will not reveal your identity. If you agree to participate in this experiment and allow us to use your results in our research, please check the consent option (attached consent form) and start the experiment in a quiet environment using headphones.

如果您在实验过程中有任何疑问或需要帮助，请随时发送电子邮件至c.lin.22@student.rug.nl。更多细节，请查看此同意书。再次感谢您的支持与参与！

If you have any questions or need assistance during the experiment, please feel free to email c.lin.22@student.rug.nl at any time. For more details, please refer to this consent form. Thank you again for your support and participation!

您同意并愿意继续参与听力实验吗?

Do you agree and are you willing to continue the listening experiment?

❑是的，我同意。(Yes, I agree.) ❑ 不，我不同意。(No, I do not agree.)

Demographic questions

您的母语是中文吗? (Is Chinese your native language?)

❑ 是的(Yes) ❑ 不是(No)

您的性别是? (What is your gender?)

❑ 男(Male) ❑ 女(female)

您是否具有与语音技术相关的背景？例如，您是否学习过语音识别或语音合成相关专业，或者在这个领域有工作经验，或者对这方面有深入的了解？(Do you have a background in speech technology? For example, have you studied majors related to speech recognition or speech synthesis, worked in this field, or have in-depth knowledge in this area?)

❑ 是的(Yes) ❑ 不是(No)

Audio instructions:

这里有3个合成的音频文件，您可以利用它们来调整音量至适中水平，并熟悉生成音频的语音特点。在之后的实验中，请不要调整手机或电脑的音量和音频的播放速度！

Here are three synthesized audio clips that you can use to adjust the volume to a moderate level and familiarize yourself with the characteristics of the synthetic speech. Please do not adjust the volume or playback speed of your phone or computer during the experiment!

Test question:

下面是一个测试，旨在让您熟悉本测试的题型。您的答案不会被记录。

The following is a test question designed to familiarize you with the types of questions in this test. Your answers will not be recorded.

想象一下，如果有人以这样的语气与您交谈，您会觉得他是以多么友好的态度与您交流呢？（请使用以下评分标准：1星表示中立（不带有特定情感或态度），2星表示略带友好，3星表示友好，4星表示非常友好，5星表示极其友好。）

Imagine if someone spoke to you in this tone, how would you rate their friendliness towards you?(Please use the following rating criteria: 1 star represents neutral (without specific emotion or attitude), 2 stars represent slightly friendly, 3 stars represent friendly, 4 stars represent very friendly, 5 stars represent extremely friendly.)

音频1 (Audio 1): ☆☆☆☆☆

音频2 (Audio 2): ☆☆☆☆☆

音频3 (Audio 3): ☆☆☆☆☆

Main question: 10x

接下来是正式的测评内容，请认真回答下列问题。

The following is the formal evaluation content. Please answer the following questions carefully.

想象一下，如果有人以这样的语气与您交谈，您会觉得他是以多么友好的态度与您交流呢？（请使用以下评分标准：1星表示中立（不带有特定情感或态度），2星表示略带友好，3星表示友好，4星表示非常友好，5星表示极其友好。）

Imagine if someone spoke to you in this tone, how would you rate their friendliness towards you?(Please use the following rating criteria: 1 star represents neutral (without specific emotion or attitude), 2 stars represent slightly friendly, 3 stars represent friendly, 4 stars represent very friendly, 5 stars represent extremely friendly.)

音频1 (Audio 1): ☆☆☆☆☆

音频2 (Audio 2): ☆☆☆☆☆

音频3 (Audio 3): ☆☆☆☆☆

C Data analysis

C.1 Mean Accuracy in Listening Test 1

Participants	Gender	Tech_back	Accuracy_N	Accuracy_P	Accuracy_PED
1	F	Y	0.1	0.4	0.4
2	F	N	0.2	0.2	0.2
3	M	N	1.0	0.1	0.0
4	F	N	0.5	0.3	0.4
5	M	N	0.4	0.0	0.0
6	M	N	0.3	0.0	0.0
7	F	Y	0.5	0.5	0.8
8	F	Y	0.1	0.4	0.2
9	F	N	0.3	0.4	0.3
10	F	Y	0.5	0.6	0.2
11	F	N	0.2	0.1	0.2
12	M	N	0.5	0.5	0.6
13	F	N	0.9	0.0	0.0
14	F	N	0.5	0.3	0.3
15	F	N	0.0	0.3	0.1
16	F	N	0.5	0.2	0.4
17	F	Y	0.3	0.2	0.2
18	F	N	0.6	0.6	0.4
19	M	N	0.6	0.4	0.7
20	M	Y	0.1	0.0	0.0
21	F	Y	1.0	0.7	0.7
22	F	N	0.3	0.3	0.4
23	F	Y	0.4	0.3	0.1
24	F	N	0.5	0.1	0.0
Average			0.43	0.29	0.28

C.2 The distribution of responses in Listening Test 1

In the table, ”F” and ”M” in the ”gender” column refer to female and male, respectively. ”Y” and ”N” in the ”tech_back” column denote individuals with and without a background in the speech technology field, respectively. ”F,” ”N,” ”D,” and ”A” represent friendly, neutral, distant, and authoritative attitudes, respectively, in the ”Condition N,” ”Condition P,” and ”Condition PED” columns. The cells highlighted in red in these three condition columns denote accurate answers and accuracy in each condition.

Participants	**Gender**	**Tech_back**	**Condition N**				**Condition P**				**Condition PED**			
			F	N	D	A	F	N	D	A	F	N	D	A
1	F	Y	0	0.1	0	0.9	0.4	0.5	0.1	0	0.4	0.5	0.1	0
2	F	N	0	0.2	0.5	0.3	0.2	0.5	0.2	0.1	0.2	0.4	0.3	0.1
3	M	N	0	1	0	0	0.1	0.9	0	0	0	1	0	0
4	F	N	0.1	0.5	0.4	0	0.3	0.2	0	0.5	0.4	0.1	0.1	0.4
5	M	N	0.4	0.4	0.1	0.1	0	0.5	0.2	0.3	0	0.3	0.2	0.5
6	M	N	0.6	0.3	0	0.1	0	0.6	0.3	0.1	0	0.2	0.6	0.2
7	F	Y	0	0.5	0.2	0.3	0.5	0.5	0	0	0.8	0.2	0	0
8	F	Y	0.2	0.1	0.7	0	0.4	0.1	0	0.5	0.2	0.2	0	0.6
9	F	N	0.2	0.3	0.4	0.1	0.4	0.3	0.2	0.1	0.3	0.4	0.2	0.1
10	F	Y	0.3	0.5	0.2	0	0.6	0.1	0.1	0.2	0.2	0.3	0.1	0.4
11	F	N	0.8	0.2	0	0	0.1	0.6	0.3	0	0.2	0.5	0.3	0
12	M	N	0.2	0.5	0.2	0.1	0.5	0.3	0.1	0.1	0.6	0.1	0.1	0.2
13	F	N	0	0.9	0	0.1	0	0.9	0.1	0	0	0.7	0.3	0
14	F	N	0.2	0.5	0.3	0	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.2
15	F	N	0.5	0	0.3	0.2	0.3	0.2	0.3	0.2	0.1	0	0.4	0.5
16	F	N	0.1	0.5	0.3	0.1	0.2	0.5	0.1	0.2	0.4	0.3	0.1	0.2
17	F	Y	0.4	0.3	0.3	0	0.2	0.4	0.2	0.2	0.2	0.5	0.2	0.1
18	F	N	0.1	0.6	0.3	0	0.6	0.2	0.2	0	0.4	0.1	0.3	0.2
19	M	N	0.1	0.6	0.2	0.1	0.4	0.1	0	0.5	0.7	0.1	0.1	0.1
20	M	Y	0	0.1	0.9	0	0	0.7	0	0.3	0	0.7	0	0.3
21	F	Y	0	1	0	0	0.7	0.1	0	0.2	0.7	0.1	0	0.2
22	F	N	0	0.3	0.4	0.3	0.3	0.5	0.1	0.1	0.4	0.4	0.2	0
23	F	Y	0.1	0.4	0.2	0.3	0.3	0.5	0	0.2	0.1	0.3	0.3	0.3
24	F	N	0.3	0.5	0.1	0.1	0.1	0.4	0.3	0.2	0	0.3	0.4	0.3
Average			0.19	0.43	0.25	0.13	0.29	0.41	0.13	0.18	0.28	0.33	0.19	0.20

C.3 MOS in listening test 2

Participants	Gender	Tech_back	MOS_N	MOS_P	MOS_PED
1	F	N	2.7	2.6	2.5
2	F	Y	2.4	2.3	2.3
3	M	Y	4.0	3.0	3.0
4	F	Y	1.3	4.0	4.1
5	F	Y	3.8	3.2	2.8
6	M	N	1.5	1.4	1.5
7	M	N	1.3	1.4	1.3
8	F	N	1.0	1.9	2.1
9	M	N	4.3	4.5	4.7
10	F	Y	4.1	3.4	3.3
11	F	Y	3.5	3.8	1.8
12	F	Y	1.4	2.3	2.3
13	F	Y	3.3	3.0	2.8
14	F	N	3.0	3.7	3.6
15	F	Y	3.0	3.8	3.8
16	M	N	2.6	3.4	2.2
17	M	Y	3.9	2.8	3.2
18	F	N	1.4	1.3	1.3
19	F	N	1.1	1.7	1.9
20	M	N	1.8	3.1	3.2
21	F	Y	1.9	1.6	1.7
22	F	N	1.6	2.0	1.8
Average			2.50	2.74	2.60

D Stimuli

The stimuli utilized in this research are available on a Google Drive, accessible via the following link: https://drive.google.com/drive/folders/1uepTf1ZZL2z4VhbsaOhRqVT8PGXjsGtK?usp=sharing. The folders are labeled as N, P, and PED, representing stimuli synthesized under conditions N, P, and PED, respectively. Within each folder, the audio files are named based on the structure: SpeakerNumber_condition_sentence. Additionally, there is a folder named "volume control" containing audio files used for participants to adjust volumes and familiarize themselves with synthetic audios before formal listening tests.