Does musical experience facilitate phonetic accommodation during human-robot interaction?

Yitian Hong[a], Si Chen[a,b,c,d], Han Jiang[a]

[a]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, [b]Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Hong Kong SAR, [c]PolyU-PekingU Research Centre on Chinese Linguistics, The Hong Kong Polytechnic University, Hong Kong SAR, [d]Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong SAR

Conflict of Interest Statement: The authors have declared that no competing interests existed at the time of publication.

Correspondence to Si Chen: sarah.chen@polyu.edu.hk

Abstract

Purpose: This study investigated the effect of musical training on phonetic accommodation in a second language after interacting with a social robot, exploring the motivations and reasons behind their accommodation strategies. .

Methods: Fifteen L2 English speakers with long-term musical training experience (musician group) and fifteen speakers without musical training experience (non-musician group) were recruited to complete four conversational tasks with the social robot Furhat. Their production of a list of keywords and carrier sentences was collected before and after conversations and used to quantify their phonetic accommodations. The spectral cues and prosodic cues of the production were extracted and analyzed.

Results: Both groups showed similar convergence patterns, but different divergence patterns. Specifically, the musician group showed divergence from the robot's production on more prosodic cues (mean f0 and duration) than the non-musician group. Both groups converged their vowel formants towards the robot without group differences.

Conclusions: The findings reflect individuals' assessment of the robot's speech characteristics and their efforts to enhance communication efficiency, which might indicate a special speech register used for addressing robot. The finding is more noticeable in the musician group compared to the non-musician group. We proposed two possible explanations of the effect of musical training on phonetic accommodations: one involves the training of auditory attention and working memory, and the other relates to the refinement of phonetic talent in second language acquisition, contributing to theories on the relationship between music and language. This study also has implications for applying musical training to speech communication training in clinical populations and for designing social robots to better serve as speech therapy partners.

Keywords: phonetic accommodation, musical training, human-robot interaction, L2 speakers

Introduction

Whether the experience of processing music influences individuals' speech processing has long been explored perhaps because these two processes share similar processing elements (e.g., pitch, duration, intensity) (Sares et al., 2018) and serve common communicative functions such as expressing emotions (Heffner & Slevc, 2015). In addition, they involve common cognitive mechanism, such as auditory working memory and auditory attention (e.g., Patel, 2011, 2012, 2014). Most studies have found that musical training experience benefits individuals' perception of more fine-grained phonetic details (e.g., Chen et al., 2020; Schön et al., 2004; Wu et al., 2015). However, few studies have investigated this effect on speech production, particularly in terms of applying the benefits of musical training to speech communication.

Phonetic accommodation describes a phenomenon in which speakers adjust their phonetic features according to those of their conversation partners (Lewandowski & Jilka, 2019). This adjustment has been observed not only during interactions but also persisting afterward (Hogstrom et al., 2018), providing a platform for examining the transfer effect of musical training on speech communication. The current study aimed to investigate this phenomenon by eliciting spontaneous speech from individuals with and without musical training experience. Moreover, because musical training experience has been found to fine-tune cognitive processes such as auditory working memory capacity and auditory attention, which are more demanding for second language speakers in processing the non-native speech, we recruited second language speakers as target participants to reveal a more robust effect. Instead of focusing on online accommodation during interaction, this study investigates the post-interaction maintenance of accommodation, which is more likely influenced by auditory memory and thus more susceptible to the effects of musical training. Additionally, we employed an innovative human-robot interaction (HRI) paradigm to investigate the effect of musical training on phonetic accommodation. This approach offers advantages on: controlled and consistent speech features of the conversation partner, relatively natural spontaneous speech, and insights into how humans adapt their speech when interacting with technology—an increasingly common occurrence in this digital age.

Robust Yet Conditional Link Between Musical Training and Speech Processing and Its Predictions for Speech Production

A positive effect of musical training on language processing has been demonstrated, in particular the processing of two perceptual attributes—pitch and duration. They are shared acoustic cues for processing melody and rhythm in music, and for generating linguistic meanings in speech (Chobert & Besson, 2013). Musical experience enhances pitch processing in sentence intonation and lexical tones. Musicians are more accurate in identifying pitch in sentences and pure tone sequences than non-musicians (Sares et al., 2018). They are faster and more accurate in detecting slight incongruous pitch element of sentences produced in their native (Schön et al., 2004) and non-native languages (Marques et al., 2007). Studies on tonal languages found musicians more sensitive to tonal information in Mandarin (Wu et al., 2015) and better at normalizing pitch variabilities in Cantonese (Zhang et al., 2023). On the other hand, individuals with musical training experience show advantages in processing temporal information in speech, such as subtle modulations of French syllable duration (Marie et al., 2011) Mandarin tone categorization (Chen et al., 2020), and more adept at identifying changes in temporal structure and pause in sentences (Sares et al., 2018).

Theoretical frameworks have emerged to account for the relationship between musical training and speech processing. One such framework is the OPERA hypothesis, proposed by Patel (2011, 2012 , 2014), which posits that musical training can improve the neural representation of speech under five specific conditions. These conditions include: Overlap – the brain network for processing music and speech cues should overlap; Precision – music requires higher precision in sensory or cognitive process (Patel, 2014); Emotion, Repetition and Attention –musical training should elicit strong positive emotions, incorporate frequent repetition, and require focused attention, which are all common characteristics of musical training. Supporting this hypothesis, Tierney & Kraus (2013) found a correlation between synchronization with a beat and auditory brainstem response, suggesting shared perception of timing details.

The empirical evidence above, together with theoretical frameworks, suggests a robust yet conditional link between musical training and speech processing. The relationship between speech production and perception has been confirmed by previous studies (for a review, see Diehl et al., 2004). Given musicians' enhanced ability to discern fine-grained auditory details such as pitch and duration, it is logical to assume that their production may be influenced by their heightened perceptual acuity.

A few studies have examined the effect of musical training on speech production. For example, individuals with extensive musical training tend to produce more precise harmonics series in speech and singing (Stegemöller et al., 2008). In a study on Cantonese merging tone pairs, musicians are found to be more accurate and quicker in tone identification and discrimination while their production of the merging tone pairs did not differ significantly from non-musicians (Ong et al., 2020). Although we are still unable to draw a solid conclusion, the relationship between speech production and perception, along with the conditions raised in OPERA hypothesis, might suggest that the effect of musical training on speech production is likely to exist under certain conditions. Specifically, only cues that require overlapping processing areas in music and speech, such as pitch and duration, will exhibit the musical effect on speech production. This assumption is supported by previous studies such as Pei et al. (2016), where individuals with high music aptitude tend to produce overall prosodic features in foreign language more accurately, but not vowels and consonants. However, some studies refute this selectivity, finding that individuals with high musical aptitude have more accurate pronunciation of foreign language phonemes than those with low musical aptitude (Milovanov et al., 2010). Additionally, musical hearing skills improve the effect of accent training on L2 vowel production (Jekiel & Malarski, 2021). Therefore, there exists a gap in understanding the effect of musical training on speech production, which the present study aims to address.

While the above studies have constrained the impact of musical training on articulation precision, the capacity to manipulate speech cues not only correlates with improved pronunciation accuracy but also contributes to smoother communication. The present study delves into a phenomenon emerging from speech interaction—phonetic accommodation—to explore the influence of musical training on individuals' modulation of phonetic cues during communication and the maintenance of this effect.

Shared Cognitive Mechanisms between Phonetic Accommodation and Musical Training

Phonetic accommodation refers to the phenomenon where speakers adjust their phonetic features based on those of their conversation partners during interactions. The adjustment can involve getting closer to the partner's speech, referred to as 'convergence', or distancing one's own speech features from their partners, referred to as 'divergence' (Giles et al., 1973). Individual variation in accommodation has been modelled as a result of differences in external factors influencing social motivations and internal psychological factors (Lewandowski & Jilka, 2019). More specifically, external factors, such as speakers' evaluation of interactions, are associated with individuals' motivations to manipulate social distance or enhance social communication. Regarding internal factors, individual differences in phonetic accommodation have been linked to the language talent of speakers (Lewandowski, 2012), their personality traits (Yu et al., 2013), among others.

Furthermore, Lewandowski (2012) and Lewandowski & Jilka (2019) propose that the external and internal factors mediate phonetic accommodation by impacting cognitive mechanisms related to attention and working memory. Within an exemplar-based theoretical framework (Goldinger, 1998), phonetic accommodation begins with an acquisition stage where speakers acquire the exemplars with detailed phonetic features from their interlocutors. This stage involves a 'noticing-recognition-coding' procedure (Pierrehumbert, 2006). Variations in attention resources allocation may lead to differences in perceiving, storing phonetic details, and in retrieving the appropriate exemplars during the output stage. Moreover, phonetic accommodation depends heavily on the working memory resources that allocated to it (Heath, 2017). Working memory capacity can determine the capacity for processing received speech and the details to be stored. The significant roles of attention and working memory help explain various external and internal factors, such as conversation success (Michalsky et al., 2018), where individuals are more likely to be attracted to and allocate more attention to successful conversations, and personality traits (Yu et al., 2013), where individuals who are more open to experience may signal higher engagement, potentially leading to greater attention to the interlocutor’s speech.

Hence, it is reasonable to expect that processes beneficial to working memory capacity and attention should also facilitate phonetic accommodations. Previous literature has demonstrated that musical training significantly improves individuals’ abilities in these two areas. For example, musicians are significantly faster in directing auditory attention (Strait et al., 2010), and have increased abilities to allocate their attention resources to target information (Marie et al., 2011). In terms of working memory, earlier studies such as Chan et al. (1998) have revealed that musicians develop better verbal working memory due to enhanced cognitive function in the left temporal area and they have longer auditory sequency memory spans (Tierney et al., 2008). It can be predicted that musical training will improve one’s working memory capacity and attention allocation, thereby facilitating phonetic accommodation.

Potential Effect of Musical Training on Phonetic Accommodation of Second Language Speakers

A significantly larger effect of musical training experience on non-native speech perception than on native speech perception has been reported (Jansen et al., 2023). Some tone processing studies even report benefits only in non-native contexts. For example, musical training aided non-native speakers in Cantonese tone discrimination (Mok & Zuo, 2012) and improved categorical perception of Mandarin pitch contour in non-native speakers (Chen et al., 2020), but these benefits were not found in native speakers. These discrepancies are attributed to the ceiling effect in native speakers.

Additionally, another possible explanation is that the effect of musical training is amplified when speech processing demands more attention and other cognitive resources (Dittinger et al., 2018). Non-native speech processing is one such case, supported by a study where subjects showed greater computational demands in processing L2 sentence prosody than L1 (Gandour et al., 2007). Some non-native speech perception accounts, such as the Perceptual Assimilation Model (So & Best, 2010, 2014), suggest that sensitivity and attention to fine-grained phonetic details might facilitate second language perception. It can be referred that musical training enhances speakers' perception of fine-grained sound details, thereby improving their non-native speech perception (Jansen et al., 2023).

However, whether this larger effect extends to second language speech production, specifically speech accommodation in our study, is unknown. According to the Speech Learning Model, perception accuracy is highly correlated with the establishment of phonetic contrasts in non-native sounds, which is related to the accuracy of producing non-native sounds (Flege, 1995). Therefore, we also expect a larger effect of musical training on second language speech production.

As mentioned above, phonetic accommodation involves three major processes: perceiving phonetic features of the target sound, encoding and storing the related properties, and reproducing the stimuli (Kim & Clayards, 2019). This relationship between perception and production during interaction suggests that phonetic accommodation in a second language is likely to similarly benefit from musical training.

Current Study: Investigating Phonetic Accommodation in Human-Robot Interaction

Phonetic accommodation involves mutual adjustment by both communicators. Because a common approach to measure phonetic accommodation is by assessing differences in target phonetic features between interlocutors over time, it is challenging to discern each participant's specific contributions. For example, Lehnert-LeHouillier et al. (2020) discovered that the observed accommodation in their target participants, indicated by differences between them and the experimenters, was actually influenced by the experimenters. Moreover, phonetic accommodation often shows relatively small effect sizes (Lewandowski & Jilka, 2019), making it crucial to use a controlled conversation partner to uncover a more robust effect.

A social robot presents such a benefit. It produces controlled speech devoid of phonetic accommodation and maintains consistent social complexities, which can elucidate the effect of musical training on phonetic accommodation. Yet, it remains unclear whether speakers will accommodate phonetically towards the robot as they do with humans. According to human-technology equivalence theories, such as 'Computer As Social Actor' account (CASA; Nass et al., 1994), humans tend to attribute social stereotypes to computers. It makes sense to expect that humans will transfer their accommodation strategies from human-human interaction (HHI) to HRI. On the other hand, a growing body of research suggests that humans develop a specific speech register tailored for interacting with technology (e.g., Cohn et al., 2022, 2024). Based this, we hypothesize that second language speakers in our study are also able to accommodate their speech in HRI, either by transferring their behaviors from HHI or by adopting a specific technology-directed speech register.

Nevertheless, previous studies on phonetic accommodation of human technology interaction have focused on different design or features of the robot, such as the robot's personas (Oviatt et al., 2004), manipulating the robot to entrain its prosody towards humans (Molenaar et al., 2021), the robot's speech rate (Shimada & Kanda, 2012), and the robot's voice quality (Gessinger, 2022). Other studies have examined individual variabilities in the evaluation of the robot (e.g., Cohn et al., 2020; Hong et al., 2023). They tend to investigate phonetic accommodation from a social perspective, but seldom from a phonetic perspective. The current study, based on the potential relationship between musical training and individuals' sensitivity to certain phonetic features, intends to examine whether this experience affect individuals' phonetic accommodation. Since one of the core benefits of musical training is the fine-tuning of auditory memory, investigating the maintenance of phonetic accommodation after interaction is more relevant for demonstrating its effect. Furthermore, while there is a robust yet conditional link between musical training and speech processing—primarily through overlapping brain areas involved in processing musical and speech cues—the relationship between musical training and speech production remains unclear. This study will also address this gap by examining the phonetic accommodation of different speech cues. Hence, our research questions are:

- Do L2 speakers show phonetic accommodation after interacting with the social robot?
- Do L2 speakers with musical training experience demonstrate different patterns of phonetic accommodation compared to L2 speakers without musical training experience?
- Does the effect of musical training only show on prosodic cues, or are spectral cues (i.e., vowel formants) also affected by it?

Methods

Data collection

The main task was adapted from the design used in Hong et al., (2023), which involved collecting production of a list of keywords and carrier sentences before and after four conversations with a social robot. In the conversations, the participants engaged with the robot to identify differences among four sets of pictures.

Stimuli

We selected the farm-themed picture set from DiapixUK tasks (Baker & Hazan, 2011), which originally comprised twelve pairs of cartoon images designed for the "spot-the-difference" game in English. Each theme contains four pairs of pictures that share similar vocabulary and depict identical keywords. Considering the setting involving scripting the robot to interact with the participants, we reduced the number of differences to five to simplify the task. These differences involve alternations to items (e.g., two white sheep in picture A vs. two grey sheep in picture B) or missing items (e.g., some red peppers on the floor in picture A vs. nothing on the floor in picture B). To enhance visual clarity, areas with discrepancies between picture pairs were circled and numbered. The keywords corresponding to these differences will be used to analyze phonetic accommodation.

Participants

Fifteen participants (11 females, 4 males; mean age: 21 ± 1.71) with musical training experience ("musicians") and fifteen participants (9 females, 6 males; mean age: 22.4 ± 3.07) without musical training experience ("non-musicians") were recruited from the campus. The musician group were reported receiving formal musical training for more than five years (excluding music courses taken at schools; mean years of musical training: 8.17 ± 2.97) while the non-musician group did not report receiving any formal musical training. They were all native Cantonese speakers and acquired English as their second language. Their age of English acquisition and their English proficiency are shown in Table 1. In addition to Cantonese and English, all participants reported acquiring Mandarin as a third language, with their age of Mandarin acquisition also detailed in Table 1. We limited data collection to the first three languages participants identified as acquired, as exposure to other languages was irrelevant to the current study. Two questionnaires containing this information are missing: No. 1140 from musician group and No. 0071 from non-musician group. However, they indicated that they are second language speakers of English when they came to the experiment. For the remaining participants, their English AOA and each subscore of English proficiency between the two groups are comparable. All of them signed a written consent form and were reimbursed for participation. The whole procedure was approved by the Departmental Research Committee of the Hong Kong Polytechnic University (ethics number: HSEARS20211011005).

Experiment setting

The Furhat social robot (Al Moubayed et al., 2012) served as the conversational partner in our study. It features a physical body with a neck and a movable head, projecting a light-based face for interaction. Using the voice of an American English male generated by Amazon Polly neural TTS system, the robot's speech production was pre-scripted to respond to specific triggering words. The speech volume remained consistent across all interactions with the children.

The interaction was carried out in a soundproof booth at the Hong Kong Polytechnic University's speech and language lab. The robot was positioned on a table around 85 cm from the participants, who sat facing it. A picture used to prompt speech interaction was placed on a table between them. Speech recordings were captured at a 44100Hz sampling rate with 16-bit resolution using an Audio-Technica AT2035 microphone, connected to Praat (Boersma & Weenink, 2001) on a computer.

Procedure

Each participant recorded all the keywords before the interaction. They were given a list of keywords and were asked to produce them as naturally as possible. Each keyword was spoken in singular and plural forms, twice in isolation and once in a carrier sentence "I can see the KEYWORD(s) in the picture", which should be the most frequently produced sentences during interaction. Following the recordings, the experimenter introduced the procedure of interaction to them and asked them to practice identifying differences between a pair of pictures depicting unrelated themes.

The interaction with the social robot commenced with a "say-hello" session, where the robot greeted the participants to get them familiar with its voice. The main session involved identifying differences in four pairs of pictures, each launched individually by the experimenter in a randomized order. Each task lasted for 10-15 minutes. After the interaction, the participants recorded all the keywords again, following the same procedure as that before interaction.

Data processing

Data extraction and normalization

For keywords produced in isolation, a trained student manually segmented the vowel portions before and after the interaction using Praat (Boersma & Weenink, 2001). A Praat script was adapted to extract vowel formants at 40% of the vowel portion, through linear predictive coding (LCP) with the 'To Formants (Burg)' command. To minimize anatomical variability arising from individual differences, the vowel formant values (F1 and F2) were transformed using Equivalent Rectangular Bandwidth (ERB) via the phonR package (McCloy, 2012/2023). The mean fundamental frequency (f0) and duration of the vowel portions were extracted by the script ProsodyPro (Xu, 2013). For carrier sentences, the whole sentences were manually segmented. Mean f0, maximum f0, minimum f0 and duration were extracted for analysis.

Data analysis

To quantify accommodation, we calculated the absolute difference between each human participant's mean speech feature and the corresponding feature extracted from robot's production (referred to as HRDiff, as showed in below equation).

$$HRDiff_{feat} = |Speaker_{mean} - Robot_{mean}| \qquad (1)$$

Here, *feat* denotes the speech features (F1, F2, local mean F0, global mean F0, global F0 range, local duration and global duration) we investigate. $\text{HRDiff}_{\text{feat}}$ represents the difference in a speech feature between the speaker and the robot before and after the interaction. $\text{HRDiff}_{\text{feat}}$ after the interaction was compared with that before the interaction. As the persistence of convergence effects has been firmly reported in previous literature (e.g., Delvaux & Soquet, 2007; Pardo, 2006), in this study, we regarded the adjustments after interaction as signals of accommodation carried over from the interaction period. A significantly higher $\text{HRDiff}_{\text{feat}}$ after the interaction indicated divergence, while a significantly lower value indicated convergence.

For statistical modeling, we initiated with a basic linear mixed effect model (Bates et al., 2015) in R. $\text{HRDiff}_{\text{feat}}$ served as the response variable, with subject and item as random variables. We incrementally added 'period' (before interaction vs. after interaction), 'musicianship' (musician vs. non-musician), their interaction terms, conducting a likelihood ratio test for each addition until we obtained the optimal model. We conducted post-hoc analysis using 'emmeans' (Lenth et al., 2023) package in R. The full emmeans outputs were reported in Appendix I.

Results

Figure 1 approximately here.

Spectral cues

For $HRDiff_{F1}$, we observed significant improvement of the model by adding 'period' (Df = 1, Chisq = 9.0929, p = 0.002**) and then 'musicianship' (Df = 2, Chisq = 5.393, p = 0.025*) as fixed effects. The interaction between 'period' and 'musicianship' did not significantly improve the model (Df = 1, Chisq = 0.1174, p = 0.73). $HRDiff_{F2}$ showed similar results. Only 'period' significantly improved the model (Df = 1, Chisq = 5.424, p = 0.020*). This suggests that participants accommodated their F1 and F2 after interaction, regardless of their musical training background. Figure 1 shows that both $HRDiff_{F1}$ and $HRDiff_{F2}$ reduced after interaction, compared to the values before the interaction, indicating that the participants converged their F1 and F2 towards the robot's production after interacting with the robot.

Figure 2 approximately here.

Prosodic cues

Local cues

We investigated the mean F0 and duration of HRDiff for each keyword produced in isolation to analyze accommodation in local cues. Adding 'period' as a fixed effect significantly improved the model for mean F0 HRDiff (Df = 1, Chisq = 21.712, p =

0.000***). The interaction between ‘period’ and ‘musicianship’ did not significantly affect the model, indicating both groups accommodated mean F0 in similar manner. As shown in Figure 2 (top left), both groups enlarged HRDiff of local mean f0 after interaction, indicating divergence. Musicians diverged more, though not significantly.

For HRDiff of duration, we found a main effect of ‘period’ (Df = 1, Chisq = 94.173, p = 0.000***) and an interaction between ‘period’ and ‘musicianship’ (Df = 1, Chisq = 63.667, p = 0.000***). Post-hoc analysis showed this effect was contributed by musician group, who significantly accommodated the duration of keywords after interaction (Df = 1160, t.ratio = -12.907, p = 0.000***). Figure 2 (top right) shows musician group significantly enlarged HRDiff of duration after interaction, indicating divergence, while non-musician group tended to maintain HRDiff after interaction.

Global cues

For global mean F0 in HRDiff, adding ‘period’ (Df = 1, Chisq = 30.903, p = 0.000***) and its interaction with musicianship (Df = 2, Chisq = 23.724, p = 0.000***) significantly improved the model. Post-hoc analysis showed musicians significantly accommodated global mean F0 after interaction, (Df = 567, t.ratio = -7.536, p = 0.000***), indicating significant divergence (Figure 2, bottom left).

For global duration in HRDiff, the addition of ‘period’ (Df = 1, Chisq = 22.57, p = 0.000***) and its interaction with ‘musicianship’ (Df = 2, Chisq = 6.007, p = 0.049*) also improved the model significantly. Post-hoc analysis showed non-musicians significantly accommodated global duration after interaction (Df = 572, t.ratio = 5.14, p = 0.000***). The reduction of HRDiff in carrier sentences (Figure 2, bottom right) indicates convergence towards the robot's duration.

In addition, we analyzed F0 range of the whole carrier sentences to examine accommodation of sentence intonation. For HRDiff in F0 range of the carrier sentences, we found a main effect of 'musicianship' (Df = 1, Chisq = 4.55, p = 0.033*) and a marginal interaction effect between 'period' and 'musicianship' (Df = 2, Chisq = 4.76, p = 0.092.). Post-hoc analysis revealed significant HRDiff differences between groups before interaction (Df = 72.2, t.ratio = 2.89, p = 0.025*), but not after (Df = 71.9, t.ratio = 0.63, p = 0.923). Figure 3 shows a trend of musicians reduced HRDiff in f0 range after interaction, while non-musicians maintained consistent HRDiff.

Figure 3 approximately here.

Discussion

The current study investigated the effect of musical training experience on L2 English speakers' maintenance of phonetic accommodation after interacting with a social robot. We analyzed spectral cues (F1 and F2) and prosodic cues (mean F0, duration, and F0 range) of keyword and carrier sentence production. **Error! Reference source not found.** summarizes the main findings. Convergence occurred when speakers' features became closer to the robot's corresponding feature after the interaction, while divergence indicated the opposite. Both groups converged spectral cues towards the robot, with no significant group differences. However, we found more evidence of accommodation in musicians than non-musicians, supporting the prediction that musical training enhances phonetic accommodation during interaction.

Figure 4 approximately here.

Possible realizations of effect of musical training on phonetic accommodations

Overall, we observed more post-interaction phonetic accommodation in musician group than non-musician group, supporting our prediction that musical training experience facilitated phonetic accommodation. We offer two potential explanations for these beneficial effects, as depicted in Figure 4. One explanation involves the training of auditory attention and working memory, while the other relates to the refinement of phonetic talent in second language acquisition.

Given that previous research has identified attention and memory as core cognitive mechanisms underlying phonetic accommodation (Lewandowski & Jilka, 2019). Musical training likely enhances these abilities, thereby facilitating phonetic accommodation. Existing literature links musical training to auditory attention (Marie et al., 2011; Strait et al., 2010), showing that playing music trains one's ability to effectively share attentional resources in auditory domain. Studies on musical training and working memory (Chan et al., 1998; Tierney et al., 2008) attribute improved working memory to musicians' intensive experience in perceiving music, which tunes their processing of auditory temporal sequences in immediate memory. Our study suggests that these benefits extend from speech processing to production. Speakers with musical training experience are superior in directing their attention to particular auditory cues. These cues are encoded and stored in their memory with greater details, enabling them to retrieve a more complete exemplar for comparison and select a more proper exemplar for production. On the other hand, according to a previous study (Feng et al., 2021), working memory is required to process pitch details. When other tasks occupy these resources, listeners' sensitivity to pitch details is reduced. It can be inferred that musicians' enlarged auditory working memory capacity may help preserve their sensitivity to phonetic details, especially given that the tasks in the current study also placed high demands on working memory in other areas. Furthermore, because this study focuses on phonetic accommodation during post-interaction period, musicians with greater auditory working memory capacity are more likely to retain the speech details for a longer duration, which aids them in selecting appropriate production tokens even after interaction.

This idea supports the OPERA hypothesis (Patel, 2014), which posits that music benefits speech when both demonstrate overlapping cognitive processes, with music having higher demands. During music playing, incoming sounds can only be stored as acoustic details, requiring greater auditory working memory capacity (Patel, 2014). Playing music also demands higher selective auditory attention to particular dimensions of sound to ensure the music is in tune (Patel, 2014). In conversation, speakers store the sounds both in acoustic forms for preparing phonetic accommodation and in semantic forms for understanding content. Although speakers perceive and store the sound for phonetic accommodation, they do not consciously attend to particular sound dimensions. The demands of auditory working memory capacity and selective auditory attention in phonetic accommodation are not as high as music playing. Therefore, the positive effect of musical training on phonetic accommodation is well explained.

On the other hand, as proposed by Lewandoski (2019), phonetic talent plays a crucial role in predicting phonetic accommodation among second language speakers.

Those with greater phonetic talent tend to exhibit more accommodation toward a second language compared to those with less talent. The concept ‘phonetic talent’ stems from the notion of “talent for accent” in second language acquisition, indicating an innate and variable ability to acquire second language pronunciation (Dornyei & Ryan, 2015). It represents cognitive-based learner variabilities (Dornyei & Ryan, 2015). This variability suggests that phonetic talent influences both the quantity and quality of stored exemplars perceived from conversation partners, as well as one's ability to successfully access matching exemplar pools and select the most appropriate exemplars for production (Lewandowski, 2012), resulting in varying degree of accommodation. The individual variabilities of phonetic talent intersect with the cognitive mechanism of phonetic accommodation, as depicted in Figure 4.

Phonetic talent has been reported to be associated with specific growth in certain brain areas (Geschwind & Galaburda, 1985). Neurobiological evidence suggests that musical training enhances the neural encoding of speech sounds and improves auditory cortical structure and function, leading to increased auditory precision (Patel, 2012). That indicates that musical training has the potential to improve individuals’ hard-wired auditory ability, a significant component of phonetic talent. Moreover, research has shown that musical training enhances phonetic abilities in second language acquisition. For example, it improves L2 speakers’ perception and production of consonant contrasts (Slevc & Miyake, 2006), the nativeness of vowel production (Jekiel & Malarski, 2021), the ability to memorize and reproduce non-native sounds (Coumel et al., 2023), and overall performance in L2 pronunciation tests (Milovanov et al., 2010).

Although the current study did not measure speakers' phonetic talent, it is reasonable to infer that by enhancing phonetic abilities in L2 acquisition, musical training may improve their ability to manipulate phonetic variations and exhibit greater phonetic accommodation. Future study should aim to provide more empirical evidence on the three-way interaction between speakers' phonetic talent, musical training experiences and phonetic accommodation to further support this proposed mechanism.

Musical training experience does not affect all phonetic cues equally

Our findings found no group differences in spectral cues (i.e., F1 and F2) accommodation, but distinct group patterns in prosodic cues (i.e., pitch and duration). This suggests that musical training may not uniformly influence phonetic accommodation across all phonetic cues. This aligns with research showing musical training experience benefits prosodic processing, such as sentence intonation (Sares et al., 2018), syllable duration (Marie et al., 2011) and lexical tone (Wu et al., 2015). Performing a high-quality musical sequence requires exact timing to execute desired notes. In contrast, speech processing involves various cues beyond pitch and duration, such as semantic and contextual cues, which can aid comprehension. Consequently, the stringent requirement for precision in pitch and duration is alleviated, allowing for potential improvement. According to the OPERA hypothesis, the heightened demands of musical training enhance the plasticity of neurons responsible for encoding these signals, thereby benefiting processing of pitch and duration in speech (Patel, 2011). Our findings extend the beneficial impact of musical training from enhancing the processing of pitch and temporal information to improving the extraction of corresponding information for speech production.

Neural encoding of spectral shape is essential for vowel identification (Zatorre et al., 2002), involving mechanisms distinct from pitch processing. Unlike speech, which necessitates precise processing of spectral shape, music may not impose as high demands on spectral shape processing (Patel, 2012). Based on the OPERA hypothesis, the effect of musical training on spectral shape processing is thus constrained (Patel, 2012). Our findings support this view on the selective influence of musical training on speech processing. Vowel formants represent spectral energy peaks in speech (Shannon et al., 1995), and their processing and extraction involve neural encoding of spectral shape. Such encoding is not expected to be enhanced by musical training experience. Our result is consistent with previous studies where overall musical ability did not significantly predict vowel production accuracy (Ghaffarvand Mokari & Werner, 2018).

The finding that musical training's benefits are selective aligns with the Speech Learning Model, which correlates the accuracy of L2 sound production with perception (Best & Tyler, 2007; Flege, 1995). It also aligns with second language perceptual learning studies where perceptual training of particular phonetic features can be directly related to production of those features. For example, Wang et al. (2003) investigated whether the tone contrasts gained perceptually transferred to production. In their study, non-native speakers recorded a list of Mandarin words before and after perceptual training. Native-speakers' perceptual judgements revealed significant improvement of Mandarin tone contrast after training. Acoustic analyses further revealed the nature of the improvement, showing that post-training tone contours approximate native norms to a greater degree than pretraining tone contours (Wang et al., 2003). Their study demonstrated that benefits gained in perception can be directly reflected in production. Similarly, our study found consistent results, showing that the speech aspects affected by musical training are consistent between perception and production.

Accommodation strategies in human-robot interaction: technology-human-equivalence or technology-directed speech?

In this study, we observed phonetic accommodation in HRI, confirming our general hypothesis that individuals are able to accommodate their speech features when interacting with a non-human agent. However, we observed that different speech cues were realized by different strategies: both groups converged vowel formants towards the robot after interaction, while they employed a mix of convergence and divergence strategies for prosodic cues.

As established in technology equivalence accounts such as Computers As Social Actors paradigm (CASA; Nass et al., 1994): people are more likely to apply between-human social behaviors when interacting with techonology. This account is able to explain parts of the findings in the current study. In previous HRI study, speakers with prestigious accents led to increased convergence of vowel formants from the interlocutors, especially among bilingual speakers (Gnevsheva et al., 2021). In our HRI study, our social robot produced standard American English, which might be perceived as prestigious accent by second language speakers, triggering their convergence of vowel formants. Regarding divergence in prosodic cues, previous studies on HHI where reported divergence mostly associated it with subjects' evaluation of the partner. In face-to-face situations, with decreasing evaluation of likability, speakers were detected more divergence of pitch-accent realization patterns away from their conversation partners (Schweitzer et al., 2017). Similarly, Abrego-Collier et al. (2011) found that negative evaluations of narrators led to divergence in voice onset time. Speech accommodation theory posits that speakers may diverge their speech behaviors when they want to bring their interlocutor's speech patterns to "a personally acceptable level" (Beebe & Giles, 1984, p. 8). Unlike vowel formants, the robot's prosody may not be a representable standard target for second language speakers. Although participants perceived robot's vowel production as positive and prestigious, they likely evaluated the robot's prosody relatively negatively and wished to bring the robot's speech into a more acceptable level by diverging their own speech.

There are reasons why the robot's prosody is evaluated more negatively. Despite efforts to make it sound natural, the robot's voice is still synthetic. In the current study, the robot used a synthetic voice called Matthew from Amazon Polly TTS platform. A survey study (Cambre et al., 2020) reported that Matthew voice received an average score of 29 points for voice quality. This score included questions about whether the voice sounded monotone, natural, or lacked emotion/personality, which is still lower than the lowest human voice rating of 63 points. In fact, in the study of Abrego-Collier et al. (2011), although the recorded narrative was real human voice with VOT manually extended, the shadowers still noticed and described the speech as 'robotic' and 'unusual', leading to divergence. The synthetic nature of the voice influences participants' accommodation strategies, consistent with prior research on synthetic voices (Cohn et al., 2019). Similarly, Gessinger et al. (2021) reported reduced convergence and occasional divergence when shadowing synthetic voices compared to shadowing natural voice. The authors argued that participants could perceive the non-humanness of synthetic voice, which created a sense of social separation. In a conversation study, Zellou et al. (2021) observed that speakers only perceptually aligned with human interlocutors, not voice-AI interlocutors. The similarity attraction theory may explain these findings, as individuals tend to be drawn to those who are similar to themselves (Sutton et al., 2019). Because the prosody of robots is less similar to that of human participants, they might feel less attracted by it and less inclined to converge. Given the importance of prosody in conveying emotions, in language and particularly in music (Jansen et al., 2023), participants with musical training in our study may be more sensitive to the robot's unusual prosody. This sensitivity could explain their tendency to diverge more: musicians diverged the mean f0 of both keywords and carrier sentences while non-musicians diverged only keywords.

However, since we did not compare the findings between HRI and HHI in this study, the conclusion inspired by previous HHI studies should be interpreted with caution. On the other hand, speakers might accommodate to facilitate communication efficiency, as predicted by CAT (Giles et al., 1973). This also aligns with listener intelligibility accounts such as Audience Design theory (Clark & Murphy, 1982), which suggests that speakers adjust their speech features to meet the need of their interlocutors.

Figure 1 approximately here.

In our study, in addition to mean f0, the accommodation of other cues yielded inconsistent results. Musicians significantly diverged keyword duration after interaction. Upon closer analysis, this involved a significant lengthening of duration -- the participants might not intend to diverge; rather, they might aim to hyperarticulate the keywords by extending their duration during interaction, and this hyper-articulation effect was carried over to post-interaction period (See Figure 5 for the raw averages of duration). This finding aligns with previous studies where speakers lengthened the speech segments and pauses for hyper-articulation when facing miscommunication with computer systems (Oviatt et al., 1998). Such lengthening was considered an attempt to increase the comprehensibility of utterances (Branigan et al., 2010). Musicians also showed a trend of converging f0 range towards the robot, which in fact represented an increase in f0 range, as shown in Figure 3. Some studies have found that speakers hyperarticulate through increased f0 range when the computer voice make more mistakes (Burnham et al., 2010; Cohn et al., 2022). Therefore, musicians' increase in f0 range may be a strategy to improve the intelligibility of their utterances and capture the conversation partner's (i.e. the robot's) attention during interaction. These two significant manipulations were observed in musicians but were absent in non-musicians, indicating that participants with musical training experience are more capable and flexible in manipulating prosodic cues to meet the robot's needs.

In fact, the lengthening of target word duration and increasement of sentence f0 range in musician group could be characteristics of a special speech register used for addressing AI-voice. An increasing number of studies are exploring this phenomenon (e.g., Cohn et al., 2022, 2024), calling it 'AI-voice directed speech', which signals a systematic adaptation of speech register when interacting with AI-voices (see review in Cohn et al., 2022). Our study contributes empirical evidence to this line of research and further demonstrates individual variability influenced by musical training in the use of this speech register.

Conclusion

This study investigated the impact of musical training on the phonetic accommodation of second language speakers after interacting with a social robot. Results showed convergence of vowel formants without group differences and a combination of accommodation strategies concerning prosodic cues, with more accommodation observed in the musician group. The beneficial effect of musical training on phonetic accommodation was confirmed, albeit with specific cues. Musical training may either directly improve auditory attention allocation and auditory memory capacity to facilitate phonetic accommodation or fine-tune individuals' phonetic talent in second language acquisition to achieve this. The mix use of accommodation strategy may be related to individuals' evaluation of robot's speech features or the need to facilitate communication efficiency, showing potential support to a specific speech register employed to interacting with AI-voice.

This study has implications for language learning programs and speech training for populations with speech problems, suggesting that musical training could enhance their phonetic accommodation abilities, thereby improving their communication skills. Additionally, designing social robots with more human-like prosody could make them better language learning companions or speech therapy partners.

Finally, this study has several limitations that offer directions for future research. First, the study did not include data during interaction, making it impossible to investigate in-the-moment accommodation. Comparing phonetic accommodation during and after interaction could provide insights into the strength of accommodation maintenance, potentially influenced by auditory working memory, which may be affected by musical training. Second, one reviewer noted that different types of musical training might impact phonetic accommodation, particularly f0 accommodation. Since we did not account for variations in musical training types, we could not directly address this issue. Future studies should either control musical training types to yield more convincing results or explore the relationship between musical training types and pitch perception and production. In addition, as another reviewer noted, factors such as speaker gender and vowel type may introduce confounding effects on vowel formant differences, potentially influencing accommodation outcomes. However, since this study focused on general accommodation patterns, we did not analyze these factors separately. Future studies should more carefully account for these potential confounding factors by incorporating them into the study design or statistical analysis to better understand their influence.

Acknowledgement

We are grateful for the valuable comments provided by the two reviewers on this manuscript.

Funding Statement

This work was supported by the Undergraduate Research and Innovation Scheme of the Hong Kong Polytechnic University and partly supported by other grants of the university (1-ZVRT; 1-ZE0D; 1-W08C; 1-WZ02), Sin Wai Kin Foundation Limited (R-ZH5Z), the Standing Committee on Language Education and Research, Education Bureau, HKSAR (K-ZB2P) and RGC (A-PB1B).

Data Availability Statement

The data are available on request from the authors.

References

Abrego-Collier, C., Grove, J., Sonderegger, M., & Alan, C. (2011). Effects of Speaker Evaluation on Phonetic Convergence. *ICPhS*, 192–195.

Abrego-Collier, C., Grove, J., Sonderegger, M., & Yu, A. C. L. (2011). EFFECTS OF SPEAKER EVALUATION ON PHONETIC CONVERGENCE. *Hong Kong*.

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). *Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction*. 114–130. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-105606

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761–770. https://doi.org/10.3758/s13428-011-0075-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beebe, L. M., & Giles, H. (1984). *Speech-accommodation theories: A discussion in terms of second-language acquisition*. *1984*(46), 5–32. https://doi.org/10.1515/ijsl.1984.46.5

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception Commonalities and complementarities. *Language Learning and Language Teaching*, *17*, 13–34. Scopus.

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

Burnham, D. K., Joeffry, S., & Rice, L. (2010). Computer- and human-directed speech before and after correction. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology, 14-16 December 2010, Melbourne, Australia*, 13–17.

Cambre, J., Colnago, J., Maddock, J., Tsai, J., & Kaye, J. (2020). Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376789

Chan, A. S., Ho, Y.-C., & Cheung, M.-C. (1998). Music training improves verbal memory. *Nature*, *396*(6707), 128–128. https://doi.org/10.1038/24075

Chen, S., Zhu, Y., Wayland, R., & Yang, Y. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? *PLOS ONE*, *15*(5), e0232514. https://doi.org/10.1371/journal.pone.0232514

Chobert, J., & Besson, M. (2013). Musical Expertise and Second Language Learning. *Brain Sciences*, *3*(4), 923–940. https://doi.org/10.3390/brainsci3020923

Clark, H. H., & Murphy, G. L. (1982). Audience Design in Meaning and Reference. In J.-F. Le Ny & W. Kintsch (Eds.), *Advances in Psychology* (Vol. 9, pp. 287–299). North-Holland. https://doi.org/10.1016/S0166-4115(09)60059-5

Cohn, M., Barreda, S., Graf Estes, K., Yu, Z., & Zellou, G. (2024). Children and adults produce distinct technology- and human-directed speech. *Scientific Reports*, *14*(1), 15611. https://doi.org/10.1038/s41598-024-66313-5

Cohn, M., Sarian, M., Predeck, K., & Zellou, G. (2020). Individual Variation in Language Attitudes Toward Voice-AI: The Role of Listeners' Autistic-Like Traits. *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2020-1339

Cohn, M., Segedin, B. F., & Zellou, G. (2019). Imitating Siri: Socially-mediated vocal alignment to device and human voices. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, *17*.

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics*, *90*, 101123.

https://doi.org/10.1016/j.wocn.2021.101123

Coumel, M., Groß, C., Sommer-Lolei, S., & Christiner, M. (2023). The Contribution of Music Abilities and Phonetic Aptitude to L2 Accent Faking Ability. *Languages*, *8*(1), Article 1. https://doi.org/10.3390/languages8010068

Delvaux, V., & Soquet, A. (2007). Inducing imitative phonetic variation in the laboratory. *Proceedings of the 16th International Conference of Phonetic Sciences*, 369–372.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(Volume 55, 2004), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dittinger, E., D'Imperio, M., & Besson, M. (2018). Enhanced neural and behavioural processing of a nonnative phonemic contrast in professional musicians. *European Journal of Neuroscience*, *47*(12), 1504–1516. https://doi.org/10.1111/ejn.13939

Dornyei, Z., & Ryan, S. (2015). *The Psychology of the Language Learner Revisited*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/polyu-ebooks/detail.action?docID=2034011

Feng, Y., Meng, Y., Li, H., & Peng, G. (2021). Effects of Cognitive Load on the Categorical Perception of Mandarin Tones. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3794–3802. https://doi.org/10.1044/2021_JSLHR-20-00695

Flege, J. E. (1995). Second-language speech learning: Theory, findings and problems.

Speech Perception and Linguistic Experience: Issues in Cross-Language Research, 233–272.

Gandour, J., Tong, Y., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y., Li, X., & Lowe, M. (2007). Neural basis of first and second language processing of sentence-level linguistic prosody. *Human Brain Mapping*, *28*(2), 94–108. https://doi.org/10.1002/hbm.20255

Geschwind, N., & Galaburda, A. M. (1985). Cerebral Lateralization: Biological Mechanisms, Associations, and Pathology: I. A Hypothesis and a Program for Research. *Archives of Neurology*, *42*(5), 428–459. https://doi.org/10.1001/archneur.1985.04060050026008

Gessinger, I. (2022). *Phonetic accommodation of human interlocutors in the context of human-computer interaction* [doctoralThesis, Saarländische Universitäts- und Landesbibliothek]. https://doi.org/10.22028/D291-35154

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication*, *127*, 43–63. https://doi.org/10.1016/j.specom.2020.12.004

Ghaffarvand Mokari, P., & Werner, S. (2018). Perceptual Training of Second-Language Vowels: Does Musical Ability Play a Role? *Journal of Psycholinguistic Research*, *47*(1), 95–112. https://doi.org/10.1007/s10936-017-9517-8

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a Theory of Interpersonal Accommodation through Language: Some Canadian Data. *Language in Society*, *2*(2), 177–192.

Gnevsheva, K., Szakay, A., & Jansen, S. (2021). Phonetic convergence across dialect boundaries in first and second language speakers. *Journal of Phonetics*, *89*, 101110. https://doi.org/10.1016/j.wocn.2021.101110

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Heath, J. (2017). How automatic is convergence? Evidence from working memory. *Proceedings of the Linguistic Society of America*, *2*, 35:1-10. https://doi.org/10.3765/plsa.v2i0.4088

Heffner, C. C., & Slevc, L. R. (2015). Prosodic Structure as a Parallel to Musical Structure. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01962

Hogstrom, A., Theodore, R., Canfield, A., Castelluccio, B., Green, J., Irvine, C., & Eigsti, I.-M. (2018). Reduced Phonetic Convergence in Autism Spectrum Disorder. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*(0). https://escholarship.org/uc/item/7gj1d252

Hong, Y., Chen, S., Zhou, F., Chan, A., & Tang, T. (2023). Phonetic entrainment in L2 human-robot interaction: An investigation of children with and without autism spectrum disorder. *Frontiers in Psychology*, *14*. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1128976

Jansen, N., Harding, E. E., Loerts, H., Başkent, D., & Lowie, W. (2023). The relation between musical abilities and speech prosody perception: A meta-analysis. *Journal of Phonetics*, *101*, 101278. https://doi.org/10.1016/j.wocn.2023.101278

Jekiel, M., & Malarski, K. (2021). Musical Hearing and Musical Experience in Second Language English Vowel Acquisition. *Journal of Speech, Language, and Hearing Research*, *64*(5), 1666–1682. https://doi.org/10.1044/2021_JSLHR-19-00253

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. https://doi.org/10.1080/23273798.2019.1582787

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.9.0) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* [doctoralThesis]. http://elib.uni-stuttgart.de/handle/11682/2875

Lewandowski, N., & Jilka, M. (2019). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication*, *4*. https://doi.org/10.3389/fcomm.2019.00018

Marie, C., Magne, C., & Besson, M. (2011). Musicians and the Metric Structure of Words. *Journal of Cognitive Neuroscience*, *23*(2), 294–305.

https://doi.org/10.1162/jocn.2010.21413

Marques, C., Moreno, S., Luís Castro, S., & Besson, M. (2007). Musicians Detect Pitch Violation in a Foreign Language Better Than Nonmusicians: Behavioral and Electrophysiological Evidence. *Journal of Cognitive Neuroscience*, *19*(9), 1453–1463. https://doi.org/10.1162/jocn.2007.19.9.1453

McCloy, D. (2023). *Drammock/phonR* [R]. https://github.com/drammock/phonR (Original work published 2012)

Michalsky, J., Schoormann, H., & Niebuhr, O. (2018). Conversational quality is affected by and reflected in prosodic entrainment: 9th International Conference on Speech Prosody 2018. *Proceedings of the 9th International Conference on Speech Prosody 2018*, 389–392. https://doi.org/10.21437/SpeechProsody.2018-79

Milovanov, R., Pietilä, P., Tervaniemi, M., & Esquef, P. A. A. (2010). Foreign language pronunciation skills and musical aptitude: A study of Finnish adults with higher education. *Learning and Individual Differences*, *20*(1), 56–60. https://doi.org/10.1016/j.lindif.2009.11.003

Mok, P. K. P., & Zuo, D. (2012). The separation between music and speech: Evidence from the perception of Cantonese tonesa). *The Journal of the Acoustical Society of America*, *132*(4), 2711–2720. https://doi.org/10.1121/1.4747010

Molenaar, B., Soliño Fernández, B., Polimeno, A., Barakova, E., & Chen, A. (2021). Pitch It Right: Using Prosodic Entrainment to Improve Robot-Assisted Foreign Language Learning in School-Aged Children. *Multimodal Technologies and*

Interaction, *5*(12), Article 12. https://doi.org/10.3390/mti5120076

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Ong, J. H., Wong, P. C. M., & Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *The Journal of the Acoustical Society of America*, *148*(6), 3443–3454. https://doi.org/10.1121/10.0002776

Oviatt, S., Bernard, J., & Levow, G.-A. (1998). Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech*, *41*(3–4), 419–442. https://doi.org/10.1177/002383099804100409

Oviatt, S., Darves, C., & Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction*, *11*(3), 300–328. https://doi.org/10.1145/1017494.1017498

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00142

Patel, A. D. (2012). The OPERA hypothesis: Assumptions and clarifications. *Annals of the New York Academy of Sciences*, *1252*(1), 124–128.

https://doi.org/10.1111/j.1749-6632.2011.06426.x

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research*, *308*, 98–108. https://doi.org/10.1016/j.heares.2013.08.011

Pei, Z., Wu, Y., Xiang, X., & Qian, H. (2016). The Effects of Musical Aptitude and Musical Training on Phonological Production in Foreign Languages. *English Language Teaching*, *9*(6), 19–29.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *34*(4), 516–530. https://doi.org/10.1016/j.wocn.2006.06.003

Sares, A. G., Foster, N. E. V., Allen, K., & Hyde, K. L. (2018). Pitch and Time Processing in Speech and Tones: The Effects of Musical Training and Attention. *Journal of Speech, Language, and Hearing Research*, *61*(3), 496–509. https://doi.org/10.1044/2017_JSLHR-S-17-0207

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341–349. https://doi.org/10.1111/1469-8986.00172.x

Schweitzer, K., Walsh, M., & Schweitzer, A. (2017). *To See or not to See: Interlocutor Visibility and Likeability Influence Convergence in Intonation*. 919–923. https://doi.org/10.21437/Interspeech.2017-1248

Shimada, M., & Kanda, T. (2012). What is the appropriate speech rate for a communication robot? *Interaction Studies*, *13*(3), 406–433. https://doi.org/10.1075/is.13.3.05kan

Slevc, L. R., & Miyake, A. (2006). Individual Differences in Second-Language Proficiency: Does Musical Ability Matter? *Psychological Science*, *17*(8), 675–681. https://doi.org/10.1111/j.1467-9280.2006.01765.x

So, C. K., & Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

So, C. K., & Best, C. T. (2014). Phonetic Influences on English and French Listeners' Assimilation of Mandarin Tones to Native Prosodic Categories. *Studies in Second Language Acquisition*, *36*(2), 195–221. https://doi.org/10.1017/S0272263114000047

Stegemöller, E. L., Skoe, E., Nicol, T., Warrier, C. M., & Kraus, N. (2008). Music Training and Vocal Production of Speech and Song. *Music Perception*, *25*(5), 419–428. https://doi.org/10.1525/mp.2008.25.5.419

Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. *Hearing Research*, *261*(1–2), 22–29. https://doi.org/10.1016/j.heares.2009.12.021

Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300833

Tierney, A., & Kraus, N. (2013). The Ability to Move to a Beat Is Linked to the

Consistency of Neural Responses to Sound. *The Journal of Neuroscience*, *33*(38), 14981–14988. https://doi.org/10.1523/JNEUROSCI.0612-13.2013

Tierney, A. T., Bergeson-Dana, T. R., & Pisoni, D. B. (2008). Effects of Early Musical Experience on Auditory Sequence Memory. *Empirical Musicology Review : EMR*, *3*(4), 178–186.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043. https://doi.org/10.1121/1.1531176

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., & Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00436

Xu, Y. (2013). *ProsodyPro—A Tool for Large-scale Systematic Prosody Analysis* [Proceedings paper]. An Interspeech 2013 satellite event. In: Tools and Resources for the Analysis of Speech Prosody. (Pp. 7 - 10). Laboratoire Parole et Langage, France: Aix-En-Provence, France. (2013); Laboratoire Parole et Langage, France. https://discovery.ucl.ac.uk/id/eprint/1406070/

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and "Autistic" Traits. *PLOS ONE*, *8*(9), e74746. https://doi.org/10.1371/journal.pone.0074746

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

Zellou, G., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience*, *36*(10), 1298–1312. https://doi.org/10.1080/23273798.2021.1931372

Zhang, K., Tao, R., & Peng, G. (2023). The advantage of the music-enabled brain in accommodating lexical tone variabilities. *Brain and Language*, *247*, 105348. https://doi.org/10.1016/j.bandl.2023.105348

Figure 1. Mean vowel formant difference in ERB transformation form between human speakers and the robot (HRDiff) before and after interaction. Error bars indicate ±1 standard error.

Figure 2. Mean difference between human speakers and the robot (HRDiff). Error bars indicate ±1 standard error. Symbols in red indicate significant differences before and after interaction.

Figure 3. F0 range produced by the robot, musician group and non-musician group. The triangle indicates the mean f0 range of each group.

Figure 4. Possible mechanisms of musical training experience affecting phonetic accommodation.

Figure 5. Duration of keywords (local) and sentence (global) produced of two groups. The triangle indicates the mean f0 range of each group.

	English AOA	**Mandarin AOA**	**listening**	**speaking**	**reading**	**writing**
Musicians	2.57 ± 1.87	4.21 ± 1.89	4.43 ± 0.51	3.93 ± 0.83	4.63 ± 0.74	4 ± 0.68
Non-Musicians	2.57 ± 1.99	4.5 ± 3.2	4.43 ± 0.51	4.29 ± 0.47	4.57 ± 0.65	4.29 ± 0.61

Table 1. Mean age of acquisition in English and self-evaluated English skills (out of 6 points). One missing data from musician group and one missing data from non-musician group.

		Musician	**Non-musician**
Spectral cues	F1	Converge	Converge
	F2	Converge	Converge
Prosodic cues	Mean f0 (local)	Diverge	Diverge
	Duration (local)	Diverge	/
	Mean f0 (global)	Diverge	/
	Duration (global)	/	Converge
	F0 range (global)	Converge (trend)	/

Table 1. summary of main results