Speech-Like Cues and the Limits of Musicality: Lexical Tone Normalization in Mandarin across Speech, Rap, and Song Contexts

Yujia Tian, Yanyuan Ye, Mency Lu

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Speech-Like Cues and the Limits of Musicality: Lexical Tone Normalization in Mandarin across Speech, Rap, and Song Contexts

Yujia Tian[1], **Yanyuan Ye**[1], **Mency Lu**[1],
JIA Fanlu[2], Ran Tao[1]
[1]Department of Language Science and Technology,
Research Centre for Language, Cognition and Neuroscience,
The Hong Kong Polytechnic University
[2]Department of Psychology, Jinan University
yujaa.tian@connect.polyu.hk, yanyuan.ye@connect.polyu.hk
mency.lu@connect.polyu.hk, spe_jiafl@ujn.edu.cn
ran.tao@polyu.edu.hk

Abstract

Lexical tone normalization enables Mandarin speakers to maintain stable tone categories despite substantial pitch variability across speakers and communicative contexts. While previous research has established that speech contexts reliably facilitate tone normalization, non-speech and purely musical contexts do not, supporting the view that this mechanism is speech-specific. However, genres such as rap, which blend speech and musical elements, challenge this dichotomy. This study systematically examined whether rap music and related vocal contexts can induce lexical tone normalization in Mandarin listeners. Native Mandarin speakers categorized target syllables following six types of auditory contexts: natural speech, clear speech (elocution), typical rap, melodic rap, song, and cello. All vocal materials were produced by professional Mandarin-speaking rappers and pitch-matched to the targets. Results revealed that natural speech, elocution, and typical rap contexts robustly elicited tone normalization, as indicated by significant shifts in categorical boundaries and improved identification accuracy. In contrast, melodic rap and song produced only marginal effects, while the cello context had minimal impact. These findings indicate that speech-like cues—particularly prosody and articulation—are critical for tone normalization, whereas increasing musicality, especially melodic structure, can inhibit this process even when some speech-like features remain. Our results refine current models of speech perception by demonstrating that lexical tone normalization depends on the presence and prominence of linguistic cues, and that melody can impose clear boundary conditions on this perceptual adjustment.

1 Introduction

Speech perception is a fundamental cognitive process that enables listeners to extract meaningful linguistic units from continuous acoustic signals. For speakers of tonal languages such as Mandarin, this process is further complicated by the need to distinguish lexical meanings based on pitch patterns, or tones. Lexical tone normalization is a crucial perceptual mechanism that allows listeners to maintain stable tone categories despite substantial variability in pitch across different speakers and communicative contexts (Li, Chen, & Wong, 2021; Peng, 2006). This ability is essential for effective spoken communication in tonal languages, where even subtle pitch differences can alter word meaning.

Variability in speech arises from numerous sources, including anatomical differences, speaker identity, emotional state, and environmental context (Peng et al., 2012). Such inter- and intra-speaker variability can obscure phoneme boundaries and pose significant challenges for listeners. To overcome these challenges, listeners rely on contextual information to normalize and categorize phonemic units, a process known as talker normalization (Leather, 1983; Wong & Diehl, 2003). Despite significant acoustic variation across talkers, listeners can recognize words, highlighting the importance of contextual cues.

Previous research has shown that speech contexts rich in linguistic and prosodic cues facilitate robust tone normalization (Leather, 1983; Zhang, Peng, & Wang, 2012). In contrast, non-speech and purely musical contexts tend to elicit little or no normalization effect, supporting the view that tone normalization is governed by speech-specific

mechanisms (Peng et al., 2012; Tao et al., 2021). However, the boundary between speech and music is not always clear-cut. Rap music, for example, occupies a unique position at the intersection of speech and music, combining articulatory and prosodic features of spoken language with musical elements such as rhythm and, at times, melody.

In this study, "typical rap" is defined as a hybrid genre that combines rhythmic speech with musical accompaniment. According to the Oxford English Dictionary and Encyclopedia Britannica, rap is characterized by spoken or chanted rhyming lyrics over a musical backing. Our focus is on the continuum between speech and music, rather than a strict categorical distinction. By examining rap as an intermediate form, we aim to explore how speech-like and musical cues interact in tone normalization.

Theoretical perspectives on talker normalization have evolved over time. The "frame of reference" theory, originally developed for vowel perception, posits that listeners use contextual information to create a cognitive reference for interpreting speech sounds (Ladefoged & Broadbent, 1957; Nearey, 1978). This theory has been extended to lexical tone perception, where contextual pitch information provides a reference for categorizing tones (Wong & Diehl, 2003; Zhang et al., 2012). However, it remains unclear whether non-speech or hybrid contexts, such as rap, can provide an effective frame of reference for tone normalization.

Recent studies have begun to explore the influence of musical contexts on tone perception. While instrumental music generally fails to induce tone normalization (Tao & Peng, 2020; Zhang et al., 2013), the effects of vocal music, especially genres that blend speech and music, are less well understood. Rap, as a genre, is characterized by rhythmic speech delivered over a musical backing, often with minimal melodic content. This hybrid nature raises important questions about the limits of speech-specific processing in tone normalization: Can rap music, with its strong speech-like qualities, induce lexical tone normalization in Mandarin listeners? Or does the presence of musicality, even in a speech-like context, inhibit this perceptual adjustment?

Our previous research (Tian, Ye, Lu, Jia, & Tao, 2024) found that rap does not impede lexical tone normalization, whereas purely musical contexts fail to trigger this effect. This suggests the existence of a threshold beyond which musical variability becomes detrimental to language comprehension. However, the precise boundary between speech-like and musical cues, and their respective roles in tone normalization, remain unclear.

The present study aims to systematically investigate the role of rap music and related vocal contexts in lexical tone normalization among Mandarin speakers. Specifically, we seek to determine whether speech-like cues—such as prosody, articulation, and prosodic structure—are sufficient to trigger tone normalization, or whether increasing musicality, particularly melodic complexity, imposes boundary conditions that limit this effect. To this end, we constructed six types of auditory contexts: natural speech, elocution (clear speech), typical rap, melodic rap, song, and cello (instrumental). All vocal contexts were produced by professional Mandarin-speaking rappers to ensure consistency of voice quality and prosody. The contexts were carefully pitch-matched to the target stimuli, which consisted of Mandarin syllables varying along a tone continuum. Native Mandarin-speaking participants, with no exposure to other Chinese dialects, were presented with each context followed by a target syllable and asked to categorize the lexical tone.

We hypothesize that while the brain's language processing systems can accommodate some degree of musical variability, excessive variability may disrupt the normalization of lexical tones. Understanding the balance between musical variability and lexical tone normalization is crucial for advancing our knowledge of language processing. To further elucidate the neural mechanisms underlying this interaction, we recorded EEG data alongside behavioral experiments.

In summary, this study seeks to refine our understanding of lexical tone normalization by exploring the effects of a continuum of auditory contexts, ranging from speech to music, on Mandarin tone perception. By systematically varying the degree of musicality in the context, we aim to identify the boundary conditions that govern the effectiveness of speech-like cues in facilitating tone normalization.

2 Methodology

We utilized a similar experimental design and stimuli as in previous research (Tao et al., 2021 ; Tian, Ye, Lu, Jia, & Tao, 2024). Below is a

brief overview of the stimuli preparation and experimental procedure; for more detailed information, refer to (Zhang et al., 2013; Zhang et al., 2017).

2.1 Participants

A total of 21 native Mandarin speakers participated in our experiment, divided into two groups: a pre-experiment group (n = 5; 3 females; mean age = 21.6 years, SD = 2.79) and a formal experiment group (n = 16; 8 females; mean age = 21.8 years, SD = 3.3). All participants were university students or recent graduates from Northern China, ensuring a high degree of linguistic homogeneity and minimizing potential confounds from regional dialect exposure. Participants were screened to confirm that Mandarin was their sole language of daily communication, and none reported any exposure to other Chinese dialects or foreign languages that might influence tone perception.

All participants were right-handed, as assessed by the Edinburgh Handedness Inventory, and had normal or corrected-to-normal vision. Prior to the experiment, participants completed a health questionnaire to exclude those with a history of hearing impairment, tinnitus, neurological disorders, or language-related difficulties. All participants had at least a college diploma, and none had received formal musical training, which could potentially affect their sensitivity to musical cues. All provided written informed consent, and the study protocol was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

2.2 Stimuli

The experimental stimuli comprised six context conditions: natural speech, elocution (clear speech), typical rap (Rap R), melodic rap (Rap M), song, and cello. The five vocal contexts were produced by four professional Mandarin-speaking rappers (two males, two females), each with over five years of professional experience and a record of Mandarin-language performances. All speakers had released albums or singles and performed at various events, ensuring high-quality and consistent vocal production. Notably, the speakers were also fluent in Cantonese, allowing for future cross-linguistic comparisons.

All recordings were made in a soundproof room using high-quality microphones. For the speech context, speakers used a normal speaking tone and speed, maintaining a natural conversational rhythm. Elocution was characterized by more focused and forceful vocalizations. For rap contexts, speakers listened to a rap accompaniment through headphones and synchronized their performance with the background music. Typical rap (Rap R) was recorded with a background music track selected for its rhythmic complexity and clear articulation, representative of Chinese Hardcore Rap. Melodic rap (Rap M), newly introduced in this experiment, featured both rhythmic and melodic elements, with melodies specified for each speaker to ensure consistency. The song context featured Chinese singing with a clear melody, while the cello context was purely instrumental, with musical notes matching the pitch height of each syllable in the rap context.

To ensure acoustic consistency, the pitch range of each speaker was measured prior to recording, and the average pitch for each context and target was calculated and aligned across conditions. The fundamental frequency (F0) of both context and target speech was analyzed using Praat software, and targets were reselected if discrepancies exceeded 10 Hz. All vocal and instrumental contexts were normalized to an intensity of 55 dB and a duration of 1800 ms. The target syllable /i/ was selected from natural recordings of the same four speakers, with Tone 1 (high-level) and Tone 2 (mid-rising) tokens chosen based on pitch trajectory and naturalness. An 11-step continuum was created between Tone 1 and Tone 2 using the STRAIGHT morphing algorithm, with steps 1, 5, 6, 7, and 11 used in the behavioral analysis to capture the full range of perceptual ambiguity.

To further distinguish typical rap from natural speech and elocution, we conducted acoustic analyses of our stimuli using Praat. The results showed that typical rap exhibited more frequent pauses (mean pause number = 2.25) and shorter average pause durations (120.67 ms) compared to speech (pause number = 1.5, pause duration = 101.71 ms) and elocution (pause number = 1.75, pause duration = 162.79 ms). These findings indicate that rap’ s rhythmic structure is more pronounced and artistically driven, whereas pauses in speech and elocution are primarily determined by semantic and syntactic boundaries. Additionally, the word pitch range in rap (17.34 Hz) was lower than in speech (21.10 Hz) and elocution (40.02 Hz), reflecting differences in prosodic variation.

Each context was further manipulated to cre-

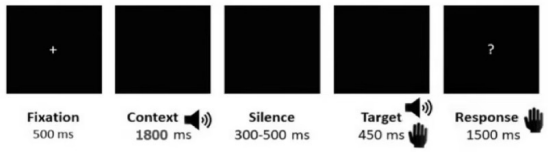

Figure 1: The trial procedure of the Mandarin word identification task.

ate two F0 conditions: F0-lowered and F0-raised, achieved by shifting the entire F0 trajectory by three semitones up or down. This manipulation allowed for the examination of contrastive context effects on tone perception. Fillers were constructed using the same procedure, with the context "接下来我会说出" ("Next I will say") and target fillers "衣" (/i55/, "clothes") and "疑" (/i35/, "misbelieve").

2.3 Experiment Procedure

Participants completed a practice block followed by six experimental blocks, each corresponding to one context condition. The order of blocks was counterbalanced across participants to control for order effects. In each trial, participants listened to a context stimulus, followed by a target syllable, and were instructed to identify the tone by pressing designated keys ("left" for Tone 1, "down" for Tone 2) with their right hand. A forward mask symbol (+) was displayed for 500 ms, followed by the context stimulus. After a brief silence (300–500 ms), the target syllable was presented. A question mark appeared after the target, prompting participants to respond within 1500 ms. Reaction times were not analyzed, as the focus was on tone identification accuracy. Each context condition included two F0 shifts (high, low), two types of content from four talkers, and five steps for the target (with the middle three steps repeated twice as often as the endpoints), plus fillers. This resulted in 72 trials per context condition. The six experimental blocks were counterbalanced to prevent order effects.

2.4 Analysis

In line with previous studies (Chen, et al., 2016; Tao, et al., 2021; Zhang et al., 2023), we analyzed the Tone 2 identification rate to assess lexical tone normalization. Our data were analyzed using Probit analysis to estimate categorical boundaries and effect sizes across contexts (Finney, 1971). The identification rate of Tone 2 was calculated for each context, F0 shift, and target step. Repeated measures ANOVAs were conducted to assess the influence of context type and F0 shift, with Greenhouse-Geisser correction applied where necessary. Post hoc comparisons were performed using Bonferroni correction.

Table 1: Derived categorical boundary positions for each type of context with high and low mean F0.

Context	High F0	Low F0	Difference
Cello	2.653	2.747	0.095
Rap M	2.942	3.050	0.107
Song	2.850	3.036	0.186
Rap R	2.618	3.209	0.591
Elocution	2.654	3.301	0.647
Speech	2.628	3.400	0.771

3 Results

The data revealed a robust hierarchy in the effectiveness of different contexts for eliciting lexical tone normalization. The speech context produced the most pronounced normalization effect, with effect sizes exceeding 20% at steps 2, 3, and 4, and peaking at nearly 30% at step 4. Elocution demonstrated a comparable, though slightly smaller, effect, with effect sizes approximately 5% lower than speech at each step and an overall mean of around 20%. Typical rap (Rap R) also induced robust normalization, particularly at step 4, where its effect size (24.11%) surpassed that of elocution (22.88%).

In contrast, melodic rap (Rap M) and song contexts produced only marginal normalization effects, with effect sizes below 10% at the middle three steps and negative values at certain steps (e.g., -0.45% for Rap M at step 4 and -2.57% for song at step 5), indicating little or no facilitation. The cello context failed to induce any significant effect, with effect sizes consistently below 5% and a mean of approximately 2.05%, suggesting minimal impact.

A 6 (context type: cello, elocution, melodic rap, rap, song, speech) Œ 2 (shift of context frequency: high, low) repeated-measures ANOVA on the categorical boundary (Greenhouse–Geisser corrected) showed no main effect of context, $F(3.26,48.86) = 1.92$, $p = .134$, $\eta_p^2 = .11$. Frequency shift, however, strongly shifted the boundary, $F(1,15) = 19.16$, $p < .001$, $\eta_p^2 = 0.56$, and this

effect was qualified by a significant interaction, $F(3.21, 48.09) = 7.83$, $p < .001$, $\eta_p^2 = .34$. Post-hoc contrasts revealed that high-frequency contexts advanced the boundary most for rap, melodic rap and song (all ps < .01), modestly for cello (p = .042), and not for speech or elocution (ps > .10).

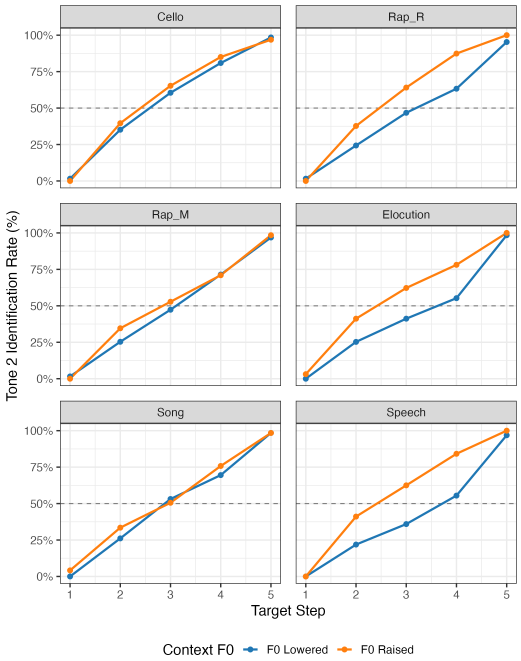

Figure 2: Average Tone 2 response by step and shift for each context (cello, melodic rap, song, rap, elocution, and speech).

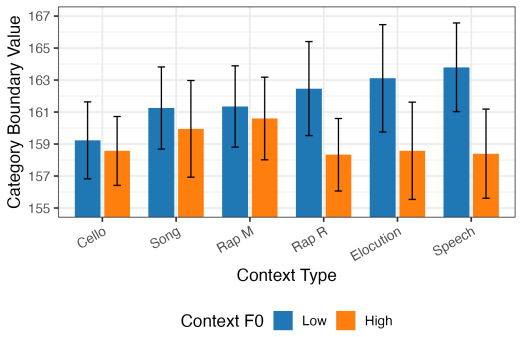

Figure 3: Category boundary values of each context type and frequency. Blue bars represent higher mean F0 contexts, whereas orange bars represent lower mean F0 contexts. Higher category boundary values indicate more Tone 2 responses.

Post hoc analyses using the LSD method were conducted to further examine differences between context types in their effects on the categorical boundary. The results indicated a significant difference between the cello context and melodic rap (Rap M) (p = .020, mean difference = -0.4054, 95% CI: [-0.7472, -0.0635]). Comparisons between the cello context and speech (p = .072), as well as the cello context and elocution (p = .088), approached significance. No other pairwise comparisons reached significance. These findings highlight the pronounced normalization effects in speech and elocution contexts, and the limited impact of melodic and instrumental contexts.

In summary, the speech context exhibited the most significant lexical tone normalization effect, with effect sizes exceeding 20% at key steps, underscoring its strong influence. Elocution, while slightly less impactful, still demonstrated substantial effects, highlighting the importance of clear linguistic cues. A typical rap context could also induce lexical tone normalization in Mandarin, particularly when pitch and rhythm align closely with speech, as seen with Rap R's effectiveness at certain steps. Conversely, Rap M and song contexts had minimal effects, likely due to melodic interference, while the cello context showed negligible impact. These results emphasize the inhibitory role of melody in tone normalization.

4 Discussion

This study provides clear evidence that speech-like cues are essential for lexical tone normalization in Mandarin. Both natural speech and elocution contexts robustly elicited normalization, as did typical rap when its rhythmic and pitch characteristics closely resembled those of speech. In contrast, contexts with greater melodic complexity, such as melodic rap and song, failed to induce significant normalization, highlighting the inhibitory role of melody.

Our acoustic analyses clarify the distinction between melody and lexical tone. By examining the mean pitch (melody) and pitch range (lexical tone variation) of each syllable across contexts, we found that rap and speech-maintained pitch contours consistent with natural Mandarin prosody. In these contexts, the pitch range within individual syllables was relatively large, indicating preserved tonal variation. However, in song and melodic rap, the pitch contours followed the imposed melody rather than natural tonal patterns, resulting in reduced pitch variation within syllables. This suggests that melody can override lexical tone cues, flattening the pitch contour and diminishing the information available for tone normalization.

To illustrate these findings, Figure 4 and Figure

5 present the mean pitch trajectories (melody) and word pitch range (lexical tone variation) for each syllable across contexts. As shown in Figure 4, the mean pitch trajectories for song and melodic rap are more stable and follow distinct melodic patterns, while speech, elocution, and typical rap display more natural pitch fluctuations that reflect Mandarin prosody. Figure 5 further demonstrates that the word pitch range is substantially reduced in song and melodic rap, indicating a flattening of tonal variation. In contrast, speech and elocution contexts maintain a much wider pitch range, and typical rap falls in between, preserving some degree of tonal variation.

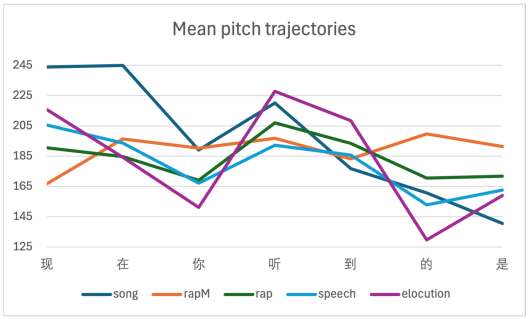

Figure 4: Mean pitch trajectories (melody) for each syllable across contexts.

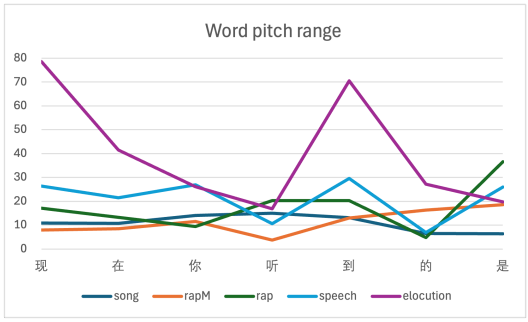

Figure 5: Word pitch range (lexical tone variation) for each syllable across contexts.

A closer look at the data reveals that, for each syllable, the pitch range in song and melodic rap remains consistently low (e.g., for “现”, song: 10.83 Hz, rapM: 7.90 Hz), while speech and elocution show much higher variability (speech: 26.35 Hz, elocution: 78.56 Hz). This pattern is consistent across all syllables, supporting the conclusion that melody in song and melodic rap overrides lexical tone cues, thereby diminishing the information available for tone normalization.

Together, these visualizations and data clearly show that as musicality increases, especially with

Table 2: Word pitch range (Hz) for each syllable across contexts.

Syllable	Song	Rap M	Rap	Speech	Elocution
现	10.83	7.90	17.04	26.35	78.56
在	10.69	8.50	13.21	21.43	41.46
你	14.00	11.52	9.45	26.93	26.06
听	15.04	3.70	20.24	10.60	16.81
到	13.17	13.04	20.21	29.51	70.48
的	6.51	16.33	4.73	6.91	27.11
是	6.35	18.54	36.53	25.95	19.68

the introduction of a strong melodic structure, both the mean pitch trajectory and the pitch range become less reflective of natural tonal variation. This supports the conclusion that melody can inhibit lexical tone normalization by reducing the availability of tonal cues in the auditory context.

These results refine our understanding of the mechanisms underlying tone normalization. While previous research has emphasized the speech-specific nature of this process (Peng et al., 2012; Tao et al., 2021), our findings suggest that the boundary between speech and music is not absolute. Instead, there appears to be a threshold of musicality—particularly the presence of a strong melodic structure—beyond which the cognitive system can no longer effectively normalize lexical tones. This threshold is most evident in the transition from typical rap to melodic rap and song.

To further clarify the effects of different context types, we conducted post hoc comparisons using the LSD method on the categorical boundary values. Although the interaction between context type and context frequency was not statistically significant, this does not undermine the main objective of our study, which was to assess the relative efficacy of each context type. The lack of interaction may be attributable to the limited sample size, which could have reduced statistical power and obscured potential effects. Some conditions exhibited clear normalization effects while others did not, suggesting that with a larger sample, more pronounced differences and interactions might emerge. Future research should therefore consider increasing the number of participants to enhance statistical power and provide a more definitive assessment of interaction effects.

Our results also contribute to ongoing debates regarding domain-general versus language-specific mechanisms in pitch processing. The lack

of normalization in purely musical contexts supports the language-specific view, while the effectiveness of speech-like rap suggests that certain musical forms can engage speech processing mechanisms under specific conditions. This aligns with the frame of reference theory, which posits that listeners use contextual cues to interpret speech sounds.

Several factors may account for the inhibitory effect of melody observed in melodic rap and song. First, a dominant melodic structure may override or interfere with the listener' s ability to use pitch information for tone normalization. Second, processing complex musical cues may increase cognitive load, reducing resources available for linguistic processing. Third, the alteration of tonal pitch information in melodic contexts may disrupt the mapping between context and target, further reducing normalization effectiveness.

Our study also highlights the importance of retaining complete tonal pitch information in the context for inducing lexical tone normalization. In song and melodic rap, although vowels and consonants are preserved, tonal cues are weakened or flattened to fit the melody, resulting in diminished normalization effects. Therefore, tonal information appears to be the most critical factor for successful lexical tone normalization.

This study has several limitations. The sample size was relatively small, and the stimuli were limited to a specific set of contexts and speakers. Future research should explore a broader range of musical genres and linguistic backgrounds and employ more detailed EEG analyses to further elucidate the neural mechanisms involved. Additionally, cross-linguistic comparisons with other tonal languages would provide valuable insights into the generalizability of these findings.

In conclusion, our results support the hypothesis that while the brain's language processing systems can accommodate some degree of musical variability, excessive musicality—particularly strong melodic structure—can disrupt the normalization of lexical tones. The threshold beyond which musical speech variability hinders language comprehension appears to lie between typical rap and melodic rap. Further research is needed to define this threshold and explore the neural mechanisms underlying this interaction.

5 Conclusion

This study demonstrates that lexical tone normalization in Mandarin is robustly supported by speech and speech-like contexts, including typical rap with strong rhythmic and articulatory features. However, as musicality increases, particularly in melodic rap and song, the normalization effect diminishes, highlighting the inhibitory role of melody. Our findings suggest that there is a threshold of musicality beyond which the cognitive system can no longer effectively normalize lexical tones.

These results refine current models of speech perception by emphasizing both the necessity of speech-specific cues and the boundary conditions imposed by musical structure. Future research should further investigate the neural mechanisms underlying these effects, explore the precise threshold between speech and music, and consider broader implications for language learning and rehabilitation.

Acknowledgments

This study has been supported by an internal grant from The Hong Kong Polytechnic University (Project No. P0051041).

References

[1] Ainsworth, W. A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, 17, 103–109.

[2] Chen, F., & Peng, G. (2016). Context effect in the categorical perception of Mandarin tones. *Journal of Signal Processing Systems*, 82(2), 253–261. doi:10.1007/s11265-015-1008-2.

[3] Dechovitz, D. (1977). Information conveyed by vowels: A confirmation. *Haskins Laboratory Status Report on Speech Research*, SR-53/54, 213–219.

[4] Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11(2), 149–175.

[5] Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America*, 125(6), 3983–3994.

[6] Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, 88, 642–654.

[7] Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), 98–104.

[8] Leather, J. (1983). Speaker normalization in the perception of lexical tone. *Journal of Phonetics*, 11, 373–382.

[9] Nearey, T. M. (1978). *Phonetic feature systems for vowels*. Indiana University Linguistics Club, Bloomington, IN.

[10] Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, 85, 2088–2113.

[11] Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, 38(4), 616–624.

[12] Peng, G., Zhang, C., Zheng, H.-Y., Minett, J. W., & Wang, W. S.-Y. (2012). The effect of intertalker variations on acoustic–perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research*, 55(2), 579–595. doi:10.1044/1092-4388(2011/11-0025).

[13] Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 40–61.

[14] Tao, R., & Peng, G. (2020). Music and speech are distinct in lexical tone normalization processing. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

[15] Tao, R., Zhang, K., & Peng, G. (2021). Music does not facilitate lexical tone normalization: A speech-specific perceptual process. *Frontiers in Psychology*, 12, 717110. doi:10.3389/fpsyg.2021.717110.

[16] Tian, Y., Ye, Y., Lu, M., Jia, F., & Tao, R. (2024). Effect of rap music context on lexical tone normalization. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 1279–1286.

[17] Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, 46(2), 413–421. doi:10.1044/1092-4388(2003/033).

[18] Ye, Y., & Peng, G. (2024). Mental representation of Mandarin Tone 3: An integrated phonetic and phonological reflection. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, 1295–1300.

[19] Zhang, C., Peng, G., & Wang, W. S.-Y. (2012). Unequal effects of speech and non-speech contexts on the perceptual normalization of Cantonese level tones. *Journal of the Acoustical Society of America*, 132(2), 1088–1099. doi:10.1121/1.4731470.

[20] Zhang, K., Sjerps, M. J., & Peng, G. (2021). Integral perception, but separate processing: The perceptual normalization of lexical tones and vowels. *Neuropsychologia*, 156, 107839. doi:10.1016/j.neuropsychologia.2021.107839

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