UNVEILING THE CONTINUITY BETWEEN MUSIC AND SPEECH: BEHAVIOURAL AND NEURAL EVIDENCE IN CANTONESE SONG PROCESSING

YE YANYUAN

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Language Science and Technology

Unveiling the Continuity between Music and Speech: Behavioural and Neural Evidence in Cantonese Song Processing

YE Yanyuan

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ _____ (Signed)

__YE Yanyuan_____ (Name of student)

Dedication

To my mother, Meijuan, whose love shaped who I am.

Abstract

The relationship between speech and music perception remains a fundamental question in cognitive neuroscience. Previous research has often framed this relationship as a binary choice between separate, modular systems and shared neural resources. This thesis challenges that dichotomy, proposing instead that the distinction between speech and music is not a single, fixed boundary but a dynamic, multi-stage process. Using the hybrid stimulus of song in Cantonese, this research investigates the interaction between music and speech domains across three progressive levels of analysis through three experiments.

The first experiment characterized the acoustic-perceptual boundary using a Categorical Perception paradigm. Listeners were presented with a seven-step acoustic continuum synthesized from speech to song across different levels of familiarity. The results did not show a classic categorical perception pattern; instead, discrimination accuracy increased continuously as stimuli became more song-like, rather than peaking at the identification boundary. This finding indicates that at the initial acoustic-perceptual level, the speech-song boundary is non-categorical, reflecting a fluid continuum where listening strategies shift gradually based on acoustic properties.

The second experiment investigated the divergence of processing at a functional level. Using a contextual normalization paradigm, the study examined how the brain's "speech mode" is selectively engaged by different vocal contexts. A clear functional hierarchy of normalization effects was observed: Speech contexts elicited the strongest effect, followed by the Song context, while Solfège and Instrumental Music contexts elicited the weakest effects. The critical dissociation between Song (sung with lyrics) and Solfège (sung with solmization syllables) demonstrates that acoustic classification

alone is insufficient to determine processing strategy. Instead, the brain's specialized speech-processing mechanisms are gated by the perceived linguistic function of the signal, placing Song in an intermediate position between pure speech and pure music.

The third experiment explored the neural correlates of this integration and the role of higher-level cognitive modulation using functional near-infrared spectroscopy (fNIRS). This study revealed a distinct neural dissociation driven by top-down factors. Familiarity with a linguistically meaningful Song led to enhanced activation in the prefrontal cortex, mirroring the neural engagement seen in speech processing. In direct contrast, familiarity with the same musical structure but without linguistic meaning (Solfège) led to reduced activation. This finding provides evidence that Song acts as a transitional state: while it shares acoustic features with music, top-down familiarity with semantic content pulls its neural processing toward the linguistic domain, overriding the efficiency mechanisms typically associated with melodic perception.

Collectively, these findings delineate a comprehensive architecture for vocal perception. Based on this evidence, a Multi-Stage Interactive Model is proposed. This framework posits that an initial, continuous analysis of acoustic features (Stage 1) is followed by a functional gating process that routes signals based on linguistic intent (Stage 2). Subsequently, these streams are not encapsulated but are subject to a final, synergistic integrative stage (Stage 3) where processing is profoundly modulated by memory and experience. This thesis concludes that the relationship between speech and music is best understood as a continuity, characterized by a dynamic interplay between bottom-up acoustic analysis and top-down functional interpretation.

Acknowledgements

First and foremost, I want to express my deepest gratitude to my supervisor, Prof. Peng Gang, for his invaluable guidance in both my research and my life. He taught me to think critically, always considering not only the feasibility but also the significance of academic research. Whenever my thoughts wandered too far afield, he was the one who guided me back on track. His dedication to experimental design and his commitment to the quality of data have influenced me a lot.

I also want to thank Prof. Wang Shi-Yuan William for offering me insightful suggestions, both in my research and in class. His dedication and passion for research have been a constant source of encouragement for me. His broad perspective and wisdom have inspired me and opened my mind.

Special thanks to Prof. Robert Zatorre, my host supervisor at McGill University. During the past six months, he has provided me with access to a supportive and enriching research environment. He encouraged me to think deeply about the theoretical contributions I could make and to pursue originality in my research, even when I felt inexperienced and uncertain. His guidance has been invaluable, and his influence will continue to shape my future directions.

I would like to thank my research fellows for their invaluable support. My heartfelt thanks go to Dr. Tao Ran, who provided inspiring ideas for my music research, and to Dr. Rong Yicheng, who guided me in conducting experiments, academic writing, and critical thinking. I am also grateful to Dr. Jérémie Ginzburg and Mr. Luo Haolun for their supports in designing and analysing the fNIRS study. This research project would not have been possible without the help of many others, including Dr. Zhang Kaile, Dr.

Weng Yi, Ms. Qi Jing, Ms. Shu Yuqin, Ms. Li Jiaxin, Ms. Ji Jinxin, Ms. Hu Yiying, Ms. Yang Yifan, Mr. Zhang Gaode, Ms. Tian Yujia, Ms. Lu Mingxi, Ms. Zhang Yu, Mr Wu Fuqian, and Mr. Zhang Junjie.

I am deeply thankful to the PolyU Choir and Mr. Alex Tam. The magical and beautiful music we created together has brought me joy, relaxation, and endless inspiration during rehearsals. I am also sincerely grateful to all the participants, whose contributions were essential to the accomplishment of this research.

I extend my heartfelt thanks to my friends. Thanks to Mr. Zeng Zhichao, for always supporting me when I'm feeling down and for keeping me grounded with his honest critiques when I lose my way. He is my true companion in both career and life. Thanks to Ms. Liu Wanqing and Ms. Song Yuhan, who have always been there for me, offering their help and company in every possible way. I am also grateful to Mr. Xie Dongyi and my cats, Migao and Guoguo, who never fail to cheer me up no matter where they are. I would also like to express my heartfelt thanks to the friends I met in Montreal and Shenzhen. Their encouragement, comfort, support, and understanding will forever be part of my cherished memories of these cities. They are D, Evan, Guo, Jennifer, Juexing, Liao, Lola, Lynn, Olivia, Sharon, and Yaoyao from Montréal, and Foher, Nanrui, Tianqi, Rou & K, and Xian from Shenzhen.

I reserve my sincere gratitude for my family, whose unwavering love and support have carried me through every challenge.

Finally, as always, I appreciate the sound of music and language, for giving me the purest enjoyment of beauty, sparking my thoughts, and serving as my eternal muse.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Research Background

The relationship between speech and music perception has long been regarded as a classic puzzle in auditory cognitive neuroscience (Aniruddh D. Patel, 2008; Zatorre et al., 2002). For decades, researchers have asked whether these two domains are processed by separate neural and cognitive systems or whether they share overlapping resources. Both faculties require the listener to extract structured and meaningful information from complex, dynamically varying acoustic signals. Fundamental mechanisms such as normalization and Categorical Perception (CP) underpin this process by enabling perceptual constancy and the mapping of continuous input onto discrete representations (Harnad, 2003; Repp, 1984). Notably, these mechanisms are observed within both music and speech, and cross-domain transfer of skills has been reported, suggesting a degree of shared processing (Bidelman et al., 2013).

However, the central debate framed as a binary choice between independent modules (Peretz & Coltheart, 2003) and shared resources (Patel, 2003) may oversimplify the nature of this relationship. A more nuanced question is rarely addressed: If specialization exists, at which processing stage(s) does it emerge? Does it occur early, at the level of acoustic-perceptual analysis, or later, during the assignment of communicative function and subsequent structural processing? Recent theories, such as the Asymmetric Sampling in Time hypothesis (AST; Poeppel, 2003) and the Shared Syntactic Integration Resource Hypothesis (SSIRH; Patel, 2003), suggest that the division between music and speech may not be a single, rigid boundary but may instead unfold dynamically across multiple processing stages.

Accumulating evidence indicates that auditory processing is inherently multi-layered (Friederici, 2002; Koelsch, 2011), and the degree of specialization may vary at different levels of analysis. For example, recent work demonstrates that spectro-temporal modulation can distinguish song from speech at the acoustic level (Albouy et al., 2024), while other findings suggest a critical divergence occurs when a signal is interpreted as linguistically functional, engaging specific mechanisms that a musically-functional signal does not (Tao et al., 2021). Finally, higher-order cognitive factors like long-term memory may further shape neural patterns at a final, integrative stage (Friederici, 2002). Thus, the critical question may not be if music and speech are distinct, but how and when their processing streams diverge or interact along a processing hierarchy from acoustic analysis to functional interpretation and cognitive integration.

Song provides a natural and ecologically valid test case for these questions. As a hybrid signal where linguistic and musical cues are simultaneously present, song challenges any simple dichotomy between speech and music, especially in tonal languages such as Cantonese. In Cantonese song, lexical tone contours are systematically mapped onto melodic structure, creating a stimulus in which linguistic and musical pitch are deeply intertwined (Chan, 1987; Kirby & Ladd, 2016; Schellenberg, 2013; Stock, 1999; Wong & Diehl, 2002). This unique stimulus is ideally suited for examining the dynamics between domain-general and domain-specific processing mechanisms, and for testing the hypothesis that the boundary between speech and music is distributed, context-dependent, and graded rather than fixed (Mehr et al., 2018, 2019; Vanden Bosch der Nederlanden et al., 2023).

In summary, this thesis moves beyond the traditional "shared versus separate" dichotomy by asking a more precise question: At which processing stage(s), and under what conditions, do music and speech diverge or converge in the brain? We propose

that specialization and overlap are not absolute, but rather emerge dynamically across multiple levels of processing, and that the perceptual and neural boundary between music and speech is fuzzy, shifting, and context dependent.

1.2 Research Objectives

Based on the multi-stage perspective outlined in the literature review, this thesis systematically investigates the separation and integration of music and speech processing across three progressive levels of analysis. The specific research objectives are as follows:

(1) To characterize the acoustic-perceptual boundary:

The first objective is to examine the fundamental nature of the perceptual space between speech and song. By employing the Categorical Perception paradigm with a seven-step acoustic continuum, this behavioural study directly tests whether the perceptual transition is discrete or continuous, thus clarifying the nature of the boundary at the acoustic-perceptual level.

(2) To investigate functional-level separation:

The second objective is to investigate how the auditory system treats vocal contexts differently based on their perceived function at a behavioural level. Using lexical tone normalization as a probe to activate speech processing strategy, this study examines whether the speech normalization effect is governed by the presence of a voice alone, or more specifically by the signal's interpretation as linguistically meaningful (as in song) versus musically meaningful (as in solfège).

(3) To explore the neural correlates of higher-level cognitive factors:

The third objective is to use functional near-infrared spectroscopy (fNIRS) to explore the neural implementation of song processing, with a specific focus on the role of

higher-level cognitive factors. This study investigates how pre-existing, long-term familiarity modulates the neural integration of lyrics and melody, testing whether this top-down factor leads to a pattern of neural efficiency or to a synergistic enhancement of brain activity.

By triangulating evidence from these three levels, this thesis aims not to reinforce a rigid division between music and speech perception, but rather to reveal the dynamic and interactive nature of auditory cognition, in which specialization, integration, and context-dependent cue weighting co-exist along the music-speech continuum.

1.3 Thesis Structure

Chapter 2 reviews the literature on share mechanisms in music and speech perception. specifically, normalization and categorical perception within music and speech domains as well as cross-domain transfer will be introduced. Additionally, the cognitive-neural models trying to figure out the relationship between music and speech processing will be reviewed, highlighting their respective strengths and limitations. Studies on song perception will also be reviewed.

Chapter 3 reports results in the CP paradigm along synthesized speech-song continua. Musicians and non-musicians from Cantonese background were recruited to complete the identification task, the discrimination task, and the familiarity judgement task. Eleven pieces of contemporary Cantonese pop songs were selected with different familiarity according to each participant. The perceptual pattern of speech-song continuum and its underlying mechanism are discussed. The effect of familiarity and musicality of participants are analysed.

Chapter 4 presents the experiments examining contextual effects of Cantonese song in lexical tone normalization paradigm among musicians and non-musicians. Solfège

(singing using solmization syllables like "*do, re, mi*"), Cantonese song, instrumental music, meaningful speech, and meaningless speech are generated as different contexts. The effects elicited by pitch shift of each context are compared.

Chapter 5 presents results of an fNIRS study exploring neural processing of song, solfège, lyrics, and typical speech among Cantonese native speakers. The effects of familiarity, vocalization type, and their interactions are discussed.

Chapter 6 summarizes main findings of three experiments, proposes a model considering both music and speech perception, and discusses implications for auditory cognition.

2 Literature review

2.1 Auditory Perception in Speech and Music: Shared Mechanisms from Distinctive Cues

Language and music represent two of the most unique productions of human, both serving as vehicles for communication. Speech and music, as acoustic signals, are characterized by high variability and thus, continuity and infinity. In speech, acoustic signals fluctuate depending on speakers' gender, vocal tract, identity, speech rate, emotional state, and even on different condition of the same individual (Appelbaum, 1996; Johnson & Sjerps, 2018). These signals are produced by the vocal apparatus and manifest as continuous changes in frequency, amplitude, and duration, theoretically spanning an infinite range of values. Similarly, music encompasses a wide array of variations, reflecting not only differences in pitch, rhythm, harmony, instruments, the performers' styles, but numerous possible combinations of pitch and rhythmic patterns that generate a rich diversity of auditory signals. This problem of *lack of invariance*, poses a fundamental challenge for the brain: to extract stable, meaningful information from constantly shifting acoustic signals.

The solution lies in a sophisticated set of cognitive processes that transform the continuous physical signals into discrete mental representations. This transformation can be conceptualized as two stages. First, the brain employs relational perception, interpreting sounds based on their relationship to one another rather than their absolute physical parameters. Second, categorization will be employed, grouping the results of this relational analysis into stable, meaningful units. These two processes represent fundamental mechanisms for processing acoustic information. The following sections will examine the operation of these mechanisms in both speech and music. This analysis

of shared and domain-specific characteristics will provide the necessary foundation for the subsequent, higher-level examination of how the cognitive systems for language and music are organized.

2.1.1 Relational and Context-Dependent Processing

Relative, instead of absolute perception, predominates in auditory cognition. Rather than primarily registering the absolute values of acoustic parameters, listeners are attuned to the relationships between sounds such as intervals and ratios. Reese (2013) states that listeners tend to rely on relative judgments when perceiving auditory stimuli, as retaining absolute values is cognitively demanding. Nonetheless, absolute perception is not entirely absent; for example, individuals with absolute pitch can identify or reproduce a given note without an external reference (Deutsch, 2013).

Pitch processing is essential in music and speech perception. As a psychoacoustic feature that is fundamentally determined by frequency, pitch is one of the eight perceptual attributes that characterize music and represents melodic contour (Pierce, 1992). In speech, pitch conveys intonation and prosody, expresses emotion, and serves as the primary acoustic cue for lexical tones in tonal languages (Gandour, 1978; Yip, 2002).

Auditory illusions

Research on auditory illusions highlights the significance of relational perception in pitch processing. The Shepard Tone Illusion (Shepard, 1964) exemplifies how listeners perceive a series of tones as endlessly ascending or descending, despite the acoustic signal cycling within a restricted frequency range. This illusion is constructed by combining multiple octave-related sine waves, resulting in a perceptual experience where the direction of pitch change is dictated by the relationship among the tones

rather than their absolute frequencies. Shepard's findings have illustrated that the perception of pitch is governed by a proximity principle within the frequency continuum, challenging the adequacy of purely linear representations of pitch.

A related phenomenon, the Tritone Paradox (Deutsch, 1986), further complicates this picture, showing that even these relational judgments are not fixed. When pairs of Shepard tones separated by a tritone are presented, listeners may perceive them as either ascending or descending, with these judgments differing both across and within individuals. Crucially, the same stimulus can elicit contrasting perceptual outcomes, as the Shepard tones indicate the pitch class with ambiguous octaves. Listeners' judgements were thus based on their internal, unique mapping between the pitch class circle and the tones they heard. This suggests that listeners rely on relational cues and contextual interpretation rather than absolute pitch information.

Both the Shepard Tone Illusion and the Tritone Paradox demonstrate that auditory perception can operate on a relational basis. When information about absolute values such as octave placement is ambiguous or cyclic, listeners default to interpreting pitch in terms of relationships and context. In other words, when absolute information is absent, the brain turns to a relational and context-dependent mode of processing. Consequently, the perception of pitch height is constructed in a context-dependent manner, and the boundaries between "higher" and "lower" are not fixed but subject to perceptual interpretation.

In general, context effect refers to a cognitive phenomenon that the behaviours and psychological responses will be alter by the surrounding information without affecting the perceptual targets or choices themselves. Context does not only function in perception, but also in attention, memory, and language (Willems & Peelen, 2021). In a micro perspective, the auditory context effect refers to the phenomenon where the

same physical stimulus can be perceived differently depending on the surrounding acoustic environment. This effect has been investigated by researchers and found robustly existing in both speech and music domains.

Context effect in music perception

The influence of context is a well-documented phenomenon in the domain of music, primarily observed in how a melodic or harmonic environment shapes the processing of single musical elements. For instance, Rakowski (1990) demonstrated that when musicians were tasked with tuning specific notes within a tonal melodic context, their precision was significantly enhanced compared to when the notes were presented in isolation. This suggested that a coherent musical framework provides critical reference for pitch judgment. Building upon this, subsequent research has shown that a global musical context, which established the key and harmonic relations of a melody, significantly influenced the local perception of musical pitch, such as interval discrimination, for both infant and adult listeners (Trainor & Trehub, 1993). This facilitative effect has been thought to arise because the melodic context provided a stable tonal reference for the target tones, thereby improving performance in pitch judgment tasks (Warrier & Zatorre, 2002).

At a neural level, ERP studies have revealed that a musical context elicited a larger mismatch negativity (MMN) amplitude in response to pitch changes compared to an isolated-tone condition, indicating that the brain processed pitch deviations more robustly in a structured context even at a pre-attentive stage (Brattico et al., 2001).

Beyond its role within the musical domain, researchers have also explored how musical context may influence aspects of language processing of songs. Wong & Diehl (2002) directly bridged musical context with lexical tone identification in Cantonese songs.

They observed that listeners apply ordinal mapping rules, where the preceding melodic contour altered the perception of the subsequent sung syllable's lexical tone. Specifically, a target lyric was more likely to be identified as having a low lexical tone when the preceding melody was high, and conversely, it was perceived as having a high lexical tone when the preceding melody was low. This finding closely mirrored the contrastive normalization effects observed in speech perception (see the next section: *Context effect in speech normalization*).

Beyond perceptual normalization, the influence of musical context also extended to higher-level cognitive functions, though its effects were more debated. The principle of context-dependent memory (CDM) suggested that memory recall was enhanced when the learning and testing contexts were congruent (Godden & Baddeley, 1975). For example, music present during a word-learning phase could facilitate subsequent recall, an effect attributed to the reinstatement of musical tonality or mood (Mead & Ball, 2007). Moreover, studies suggested this facilitative effect was particularly pronounced for the learning of infrequent words, without observing negative impacts on memory for frequent words (de Groot, 2006).

Nonetheless, the facilitative effect of musical context, particularly when used as background music, is not universally accepted and remains a subject of debate. A meta-analysis conducted by (Kämpfe et al., 2011) presented a more complex picture. Their findings indicated that while background music could have a positive impact on emotional reactions and improve performance in sports, it can also act as a distraction. Specifically, background music was found to disrupt reading comprehension and have small, yet detrimental, effects on memory tasks. This controversy underscores the need to carefully consider the nature of the musical context and the specific cognitive task being examined. The distinction between music as an integral part of the stimulus (as

in a song) versus as an external background element may be critical in determining its effect on cognitive processing.

Context effect in speech perception

In the domain of speech, the context effect is primarily understood as extrinsic normalization, a cognitive process wherein listeners use surrounding acoustic information to disambiguate sounds. To map highly variable acoustic signals onto discrete linguistic categories, the auditory system utilizes both intrinsic cues within the target sound itself (e.g., pitch height and contour for lexical tones) and extrinsic cues from the preceding or following contexts (Gandour, 1978). While early research on extrinsic normalization focused on the segmental domain of vowels and consonants (Ladefoged & Broadbent, 1957), this principle has been extended to the suprasegmental domain.

A key area of this research is lexical tone normalization. Moore and Jongman (1997) demonstrated the fundamental principle of contrastive normalization for Mandarin tones: a target syllable is perceived as having a higher tone when preceded by a lower-pitched context, and as a lower tone when preceded by a higher-pitched context. Their findings, which showed a clear identification shift for a Mandarin Tone 2-3 continuum based on context, established a foundational paradigm that has since been replicated across other Mandarin tone pairs (F. Chen & Peng, 2016; K. Zhang et al., 2018).

The study of normalization is particularly relevant for tonal languages like Cantonese, whose inventory includes three level tones (high, mid, and low). The relative acoustic similarity of these level tones means that intrinsic cues are often insufficient for disambiguation, making extrinsic context critically important for their recognition. Research has confirmed this dependency. Wong and Diehl (2003) observed that a

Cantonese mid-level tone could be perceived as a high-level tone in a lowered-pitch context or as a low-level tone in a raised-pitch context.

Subsequent research has sought to understand the specific mechanisms that drive the normalization effect. One line of inquiry has focused on the acoustic properties of the context itself. Francis et al. (2006) found that the magnitude of the normalization effect was proportional to the degree of pitch shift in the context, and that a following context could elicit a stronger effect than a preceding one. Another debate concerned the precise acoustic cues the listeners would use. Some evidence suggested listeners assessed the speaker's overall pitch range (Wong & Diehl, 2003), while other findings indicated that the mean fundamental frequency (f0) of the context was sufficient, even in stimuli with a flattened pitch contour (Francis et al., 2006).

A second, more central debate concerned the domain-specificity of lexical tone normalization. The general auditory processing account posit that normalization was a low-level acoustic process, not specific to speech processing. Support for this view came from findings that the effect could occur across different languages (Fox & Qi, 1990; Wong & Diehl, 2003) and, in one influential study, that non-speech sine-wave contexts could induce normalization for Mandarin tones (Huang & Holt, 2009).

However, the weight of the evidence has increasingly favoured a speech-specific processing account. Multiple studies have failed to replicate the non-speech effect found by Huang and Holt (2009), instead finding that contexts such as reversed speech or triangle waves did not elicit significant normalization (F. Chen & Peng, 2016). Research on Cantonese has been particularly consistent, showing that various non-speech contexts, including schwa-only sentences, triangle waves, and iterated rippled noise, produce null or negligible effects compared to speech contexts (Francis et al., 2006; C. Zhang et al., 2012; K. Zhang et al., 2018).

Instrumental music has been used as a particularly strong test case for specificity. Tao et al. (2021) found that a musical context composed of piano notes, while matched in pitch to a speech context, failed to elicit normalization effect of Cantonese level tones in either musicians or non-musicians. This provided compelling evidence that the normalization mechanism was tuned specifically to the properties of a speech signal. Interestingly, a follow-up EEG study suggested that while the behavioural outcome was the same, musicians processed the contextual cues at an earlier neural stage than non-musicians (K. Zhang et al., 2023), indicating that expertise could modulate the underlying neural dynamics even when the process remained speech specific.

Commonalities and Distinctions of Speech and Music Context Effects

A synthesis of context effects in speech and music reveals a fascinating interplay of shared auditory mechanisms and domain-specific adaptations. While both systems leverage context to resolve acoustic ambiguity, a comparison of their functions and underlying processes highlights critical distinctions and points toward a significant gap in the current literature.

At the mechanistic level, one of the most salient commonalities between speech and music perception is the principle of contrastive context effect. In both domains, the perception of a target pitch is systematically biased away from the pitch of the preceding context. This contrastive shift has been robustly observed in lexical tone perception (Moore & Jongman, 1997; Wong & Diehl, 2003), where a syllable following a high-pitched context is more likely to be perceived as lower in pitch, and vice versa. A similar effect appears in music, where melodic or harmonic context influences the perception of subsequent notes (Rakowski, 1990; Trainor & Trehub, 1993), and in song, where the melodic contour of preceding notes alters the perception of sung lexical tones

(Wong & Diehl, 2002). These phenomena suggest that the auditory system constructs a temporary reference frame based on recent input, against which subsequent stimuli are evaluated.

Furthermore, a considerable body of research demonstrates the role of experience-dependent plasticity in shaping these effects. Musicians and tonal language speakers often show enhanced processing of contextual cues, both in music and speech tasks (Bidelman et al., 2013; Rakowski, 1990; K. Zhang et al., 2023). This indicates that, while the contrastive normalization mechanism may be grounded in general auditory processing, it is also modulated by long-term domain-specific experiences.

Together, these findings support the existence of partially shared auditory processing resources, whereby the brain applies similar computational strategies to achieve perceptual constancy amidst acoustic variability in both speech and music.

Despite the shared principle of contrastive normalization, a fundamental distinction arises from the frame of reference against which sounds are judged in speech versus music. In speech perception, normalization is a process of calibrating to a specific biological source. Listeners use contextual cues to infer a talker's unique vocal characteristics, such as their vocal tract length, and use this information to achieve perceptual constancy for linguistic units like phonemes and lexical tones across highly variable speakers. In contrast, normalization in music typically operates relative to an abstract formal system. The judgement to a musical note is based on its relationship with the established key and harmonic structure of a musical piece. Therefore, music normalization is a process that is largely independent of a specific performer's identity. This distinction highlights the central role of the human voice in engaging speech-specific mechanisms. The process by which the brain extracts and utilizes this vocal

source information is therefore crucial to understanding the specialization of the speech normalization system.

The finding that vocal contexts elicit stronger normalization than non-vocal ones is best understood through the Vocal Source Hypothesis. This framework posits that speech normalization is critically dependent not just on acoustic contrast, but on the extraction of speaker-specific information from the vocal signal. This hypothesis is strongly supported by neuroimaging research. For instance, studies have identified specialized regions in the superior temporal sulcus (STS), often termed Temporal Voice Areas (TVA), that respond selectively to human vocal sounds, distinguishing them from other complex sounds like instrumental music (Belin et al., 2000). This neural specialization for the human voice provides the biological foundation for normalization mechanism.

The primary proposed mechanism for normalization process is vocal tract normalization. According to the source-filter model of speech production, the vocal tract acts as a filter to shape the sound produced by the vocal folds (the source) into distinct phonemes (Fant, 1971). Since a primary source of acoustic variability between speakers (e.g., men, women, and children) is the length and shape of their vocal tract, listeners must rapidly estimate a speaker's vocal tract characteristics to accurately perceive speech sounds (Johnson & Sjerps, 2018).

The vocal tract normalization could explain the behavioural observation that robust normalization effects are elicited only when the context is perceived as originating from a human vocal source. In contrast, instrumental music and non-speech stimuli, even when closely matched in pitch and temporal structure, generally fail to elicit the same degree of normalization (Francis et al., 2006; Tao et al., 2021). Thus, the available evidence indicates that contrastive normalization in speech is a speaker-specific,

biologically grounded process, whereas in music it is more abstract and structurally oriented.

Despite extensive research, most empirical work on normalization has contrasted speech with non-vocal contexts, such as instrumental music or non-speech synthetic sounds, while overlooking song (vocal music) as a stimulus. Song is unique in that it combines the presence of human vocal cues with a structured musical context. It carries rich information about the singer's vocal tract and identity, while simultaneously conforming to musical principles of pitch and rhythm.

From a theoretical perspective, song constitutes a stringent test case for domain-specificity in normalization mechanisms. If normalization in speech is triggered by the detection of a human vocal source, then song contexts, regardless of whether they are linguistic or musical, should elicit normalization effects comparable to those observed in speech. Alternatively, if normalization depends primarily on linguistic mode activation or speech-specific processing pathways, then song, especially when lacking linguistic semantic content or emphasizing melodic structure, may not elicit the same effects. The inclusion of song as an experimental context offers a unique opportunity to disentangle the relative contributions of vocal source information and abstract musical structure in normalization. Systematic investigation of normalization in song could clarify whether the underlying mechanism is truly speech-specific, voice-specific, or dependent on higher-order linguistic processing.

In summary, while both speech and music utilize contrastive normalization to stabilize perceptual categories, they do so with reference to fundamentally different anchors: speech relies on dynamic adaptation to individual vocal sources for linguistic identification, whereas music relies on abstract tonal frameworks for structural and aesthetic evaluation. Song, as a natural hybrid of voice and music, represents a critical

test case for the field. Its dual status allows us to directly assess the boundaries and interactions of speech and music perception. Research on normalization in song contexts is therefore essential for advancing theoretical understanding of domain specificity and shared mechanisms in auditory perception.

2.1.2 Categorization and Categorical Perception (CP)

As the preceding section has demonstrated, the brain navigates the challenge of acoustic variability by employing relative perception and leveraging contextual information. This process enables listeners to interpret sounds dynamically based on their surrounding environment. However, this relational analysis alone is insufficient to form the stable mental objects required for speech and music. An interpretation of relationships does not automatically yield the discrete categories, such as phonemes, lexical tones, or musical notes.

To transform the continuous, context-sensitive output of relational processing into the stable, meaningful units needed for higher-level comprehension, the brain needs to perform a second, decisive operation: categorization. A powerful and extensively studied manifestation of this process is Categorical Perception (CP), a mechanism that segments a continuous physical signal into discrete mental representations. This section will define the principles of CP, outline its classic experimental paradigm, and detail its foundational role in auditory perception, where it provides the stable phonological or musical units upon which speech and music is built.

CP refers to a specific perceptual phenomenon where a continuous change in a physical stimulus is perceived as discrete, bounded categories. The hallmark of CP is a "perceptual warping" effect (Harnad, 2003), characterized by two key features: between-category separation, where discrimination between stimuli from different

categories is sharp and accurate, and within-category compression, where discrimination between stimuli that fall within the same category is poor, often near chance level (Repp, 1984). This cognitive mechanism effectively highlights linguistically relevant differences while suppressing irrelevant acoustic variations. While CP has been considered as a domain-general cognitive function, observed in the perception of colours, facial expressions, and music (Etcoff & Magee, 1992; Goldstone & Hendrickson, 2010), its canonical and most robust demonstrations were found in speech perception.

Speech CP

The phenomenon of CP was first documented in the research conducted at Haskins Laboratories in the 1950s. Using the Pattern Playback synthesizer, researchers created a continuum of stop-consonant sounds by systematically varying the acoustic cue of the second formant transition, creating stimuli that ranged from /b/ to /d/ to /g/ (Liberman, 1957; Liberman et al., 1957). They tested listeners' perception of these stimuli using a two-task experimental paradigm that has since become classic. The first part is the identification task, where listeners are presented with each stimulus from the continuum and must label it as belonging to one of the possible categories (e.g., /b/ or /d/). The typical result is a sharp sigmoidal identification function, where responses abruptly switch from one category to the other at a specific point along the continuum, known as the categorical boundary. The 50% of the identification rate of one category is considered as the boundary position. The steepness of this function's slope reflects the strength of the categorization. The second part is the discrimination task (e.g., an AX task), where listeners hear two sounds and decide whether the second sound (X) is identical or different to the first sound (A). The key finding is a distinct peak in

discrimination accuracy for pairs of stimuli that cross the boundary identified in the first task, while discrimination for pairs within a category remains poor. The correspondence between the identification boundary and the discrimination peak provides evidence that listeners do not perceive the acoustic continuum veridically; instead, their perception is warped by the existence of categories.

The function of CP in speech perception is essential. Primarily, it serves to achieve perceptual constancy, allowing a listener to recognize the same phoneme despite enormous acoustic variations caused by different speakers, accents, speech rates, and coarticulation effects (Kuhl, 2004). By compressing within-category variation, CP allows the brain to treat physically different sounds (e.g., a /d/ spoken by a man and a /d/ spoken by a child) as functionally equivalent, thereby ensuring robust communication. This categorical processing also enhances cognitive efficiency. By converting the complex, continuous acoustic signal into the discrete categories, the brain can process speech with the speed necessary for comprehension. These stable phonemic units then serve as the fundamental building blocks for higher-level linguistic operations, such as lexical access and syntactic parsing.

Since the pioneering work of Liberman and his colleagues, the CP of speech has been extensively demonstrated across a wide range of phonetic contrasts. While classic studies focused on the temporal cue of Voice Onset Time (VOT) to distinguish voiced and voiceless stop consonants (e.g., /b/ vs. /p/), subsequent research has shown CP for vowels, fricatives, and liquids (for a review, see McMurray, 2022).

Furthermore, CP has been shown to extend beyond the segmental domain into the suprasegmental domain, most notably in the perception of lexical tones. For instance, Peng et al. (2010) investigated the perception of Cantonese and Mandarin lexical tones and found that native speakers of these languages perceived a continuous change in

pitch contour categorically. Their identification functions were significantly steeper, and their category boundary widths were narrower, compared to those of non-tonal language speakers (German). Crucially, the discrimination performance of the tonal language speakers revealed a sharp peak aligned with a linguistically defined boundary, whereas the non-tonal speakers exhibited a weaker, more linear discrimination pattern consistent with general psychoacoustic sensitivity. This finding underscores a critical aspect of CP: it is not an innate auditory mechanism but is profoundly shaped by experience-dependent plasticity, where long-term exposure to the native language shapes the listener's perceptual space to align with the specific phonological system of that language.

Music CP

Similar to the way phonemes are perceived in speech, Categorical Perception is also a fundamental mechanism in the domain of music. It refers to the mental phenomenon in which listeners perceive continuously varying acoustic information, such as pitch or duration, as a series of discrete and stable categories. This cognitive process explains how listeners identify stable, recognizable notes, intervals, and rhythms from a performance world filled with subtle acoustic variations.

Pioneering research on musical CP focused primarily on the perception of pitch and musical intervals. Siegel & Siegel (1977) constructed a continuum of musical notes ranging at a step of 20 cents and instructed listeners to label the musical notes. Discrimination performances were rated in the Magnitude Estimation Task, a more difficult version than the classical CP paradigm, in which listeners were instructed to compare and visualize the target intervals with the standard ones. They found that musicians could easily distinguish two intervals that crossed a categorical boundary

(e.g., a clear minor third versus a major third) but struggled to differentiate between two slightly different versions of an interval from within the same category (e.g., two slightly different major thirds). This phenomenon showing high between-category discrimination and poor within-category discrimination became the key evidence for the existence of CP. Shortly after, Burns & Ward (1978) reinforced these findings using a smaller step size (12.5 cents) in stimulus construction, noting that this perceptual skill was particularly pronounced in musicians, which suggested that professional musical training plays a critical role in shaping auditory categorization.

The phenomenon of CP is not limited to the dimension of pitch. Eric Clarke (1987) successfully extended the theory to the domain of rhythm. He argued that listeners tend to perceive continuous variations in durational ratios between notes as simple, integer-ratio rhythmic patterns (e.g., 2:1 or 3:1). Clarke framed this as an "ecological perspective" on perception, in which listeners actively categorize rhythmic information within a given metrical context to achieve a stable perception of musical structure. This finding was later replicated by Schulze (1989).

Cross-Domain of CP

A central theoretical question concerns whether CP reflects a domain-general auditory mechanism which serves as an inherent property of perceptual systems, or whether it is modulated by domain-specific experience, such as language background or musical training. Previous studies have demonstrated CP in music and speech modalities, suggesting a general auditory mechanism for segmenting continuous input into functionally meaningful units (Goldstone & Hendrickson, 2010). However, evidence from both behavioural and neurophysiological studies indicates that the fine-tuning and robustness of CP boundaries are strongly shaped by domain-specific exposure and

training. As reviewed above, speech CP is robustly observed in the perception of lexical tones among native speakers of tonal languages rather than non-native speakers (Peng et al., 2010). Similarly, music CP is observed in the categorization of pitches, intervals, and rhythms, particularly among professional musicians (Siegel & Siegel, 1977; Burns & Ward, 1978; Clarke, 1987). This pattern points to the influence of long-term experience on the perceptual system, a process described as domain-specific tuning.

Recent neuroimaging research has been clarifying the relationship between domain-general and domain-specific mechanisms in CP. For example, Bidelman et al. (2013) used EEG and fMRI to demonstrate that both musicians and tonal language speakers exhibit enhanced categorical perception for pitch, with shared neural correlates in auditory cortex. Their results indicate partially overlapping neural networks for processing categorical pitch information in both speech and music, supporting the idea of a common auditory substrate that is modulated by domain-specific training. At the same time, distinct patterns of activation related to linguistic or musical context suggest that domain-specific adaptation refines the boundaries and efficiency of CP in each domain.

Moreover, cross-domain transfer effects have been documented. Musical training can enhance CP for lexical tones in non-tonal language speakers (Chen et al., 2020), while tone language background can facilitate musical pitch categorization. Aiming to examine the effect of musical training in lexical tone CP in tonal and non-tonal language, Chen et al. (2020) recruited English musicians and Chinese musicians as well as the pairing non-musicians to participate in identification and discrimination tasks in the CP paradigm. The continua were carefully constructed considering multiple factors: duration, intrinsic f0, and pitch direction. Their results showed that musicianship has enhanced the degree of CP for non-tonal native speakers but not for tonal language

speakers. Additionally, musicians could make better use of duration and be compensated more for intrinsic f0. These results have indicated a cross-domain transfer from music to speech in the CP framework. These effects are evident not only behaviourally but also in neural signatures, such as enhanced mismatch negativity (MMN) and N1 responses during categorical discrimination tasks (Bidelman et al., 2013; Zatorre & Gandour, 2007).

In summary, while CP can be described as a domain-general perceptual strategy for discretizing continuous acoustic input, its precise implementation is highly sensitive to domain-specific experience. Neural evidence from brain imaging and electrophysiology indicates both shared and specialized pathways for categorical processing in speech and music, and cross-domain transfer underscores the plasticity of these mechanisms. The boundaries of CP are not determined purely by auditory mechanics but are shaped by linguistic and musical experience, highlighting the dynamic and context-dependent nature of perceptual categorization in the human brain.

Commonalities and Distinctions of Speech and Music CP

While CP operates in both speech and music, a direct comparison reveals crucial commonalities and distinctions, pointing toward significant gaps in our current understanding.

The fundamental mechanism of CP is shared across both domains. It transforms the continuous, ambiguous acoustic signal into discrete, meaningful categories. This process enhances perceptual constancy, allowing for efficient information processing. Both domains are heavily shaped by experience-dependent plasticity. Just as exposure to a native language sharpens phonemic boundaries, long-term musical training refines the perception of musical categories (Burns & Ward, 1978). The work by Bidelman et

al. (2013) provides the most direct evidence of this shared principle, showing that expertise in either domain (music or tonal language) enhances the underlying perceptual abilities for pitch. This suggests that speech and music may rely on, at least partially, shared neural resources for processing acoustic features like pitch, as supported by neuroimaging studies (Zatorre & Gandour, 2007).

Despite these similarities, the function and nature of the categories differ significantly. The primary function of CP in speech is communicative and referential. Distinguishing /b/ from /p/ alters the meaning of a word (e.g., "*bat*" vs. "*pat*"). In contrast, the function of CP in music is primarily aesthetic and structural. The distinction between a C and a C# defines melodic and harmonic relationships, creating structure and emotional affect, but it does not carry a fixed semantic meaning in the same way a phoneme does.

Additionally, musical categories appear to be more flexible than phonemic categories. A musician can be trained to perceive and produce microtonal intervals that lie outside of standard systems, effectively altering their categorical boundaries. While second-language learners can acquire new phonemic categories, native phoneme boundaries are found to be relatively stable in adults (Feng, 2022; Peng et al., 2010). Furthermore, musical categories are culturally constructed (e.g., the 12-tone Western scale vs. the 22-shruti system in Indian music), whereas phonemic categories are language-specific inventions.

The complexity of acoustic cues also differs between speech and music CP. The acoustic cues for phonemic categories are often complex and multi-dimensional (e.g., voice onset time, formant transitions, and duration). In music, the primary acoustic cue for pitch categories is more singular: the f0. While other features like timbre can influence pitch perception, the relationship is arguably more direct than in speech.

2.1.3 Summary: Shared Mechanisms and Domain-Specific Functions

The review of context-dependent normalization and Categorical Perception reveals two consistent findings. First, speech and music processing appear to utilize shared underlying mechanisms. Both domains rely on contrastive normalization and Categorical Perception to create stable categories from variable acoustic signals, suggesting a common foundation for processing acoustic relationships.

Despite this shared foundation, these mechanisms are applied in domain-specific ways to serve different higher-level functions. For instance, normalization in speech is highly sensitive to vocal source cues and phonological information to ensure linguistic comprehension, whereas context effects in music primarily serve structural and aesthetic purposes. Likewise, the linguistic categories are referential in speech, defining word meaning, but musical categories are structural and affective, defining tonality and emotion.

This discrepancy between shared mechanisms and domain-specific functions raises a fundamental question regarding the overall cognitive architecture. To address this, the argument must shift its level of analysis: from examining specific perceptual mechanisms (i.e., how processes like normalization and categorization operate on an acoustic signal) to investigating system-wide cognitive organization (i.e., how the faculties for speech and music are structured in relation to each other in the brain).

This shift is logically necessary because the existence of shared mechanisms, as detailed in this section, is not in itself decisive. The findings reviewed from *Section 2.1* could be interpreted in two competing ways. From a modularity perspective, one could argue that speech and music are independent, functionally specialized systems that have simply, through convergent evolution or computational necessity, adopted similar efficient solutions for processing sound. From a shared-resource perspective, one would

argue that these shared mechanisms are direct evidence of a common, underlying processor that is flexibly utilized by both speech and music.

Therefore, to resolve the ambiguity left by the analysis of perceptual mechanisms, it is essential to examine the broader behavioural and neural evidence that directly tests the architectural relationship between the speech and music faculties. The following section undertakes this review, focusing on the long-standing debate between modularity and shared resources in speech and music perception.

2.2 The Cognitive and Neural Architecture: Stages and Mechanisms of Separation

The question of whether, and at which stage, music and speech become functionally separated in the brain has shaped decades of theoretical and empirical work. Here, I will review and critically evaluate five influential frameworks: modularity, the Asymmetric Sampling in Time (AST) hypothesis, the Shared Syntactic Integration Resource Hypothesis (SSIRH), Friederici's three-phase model on sentence processing, and Koelsch's comprehensive model on music perception, in an order that reflects both the historical development of the field and the increasing complexity of proposed mechanisms.

2.2.1 Modularity: System-Level Separation

The classic argument for a fundamental separation between speech and music processing is rooted in the modularity hypothesis, which posits that the human mind is composed of distinct, functionally specialized cognitive systems. In the domain of music, one of the most influential and detailed articulations of this view is the neurocognitive model proposed by Isabelle Peretz and her colleagues (Peretz & Coltheart, 2003). This framework, grounded in decades of neuropsychological research,

argues for the existence of a music-specific faculty that is functionally independent from the language faculty and other cognitive systems.

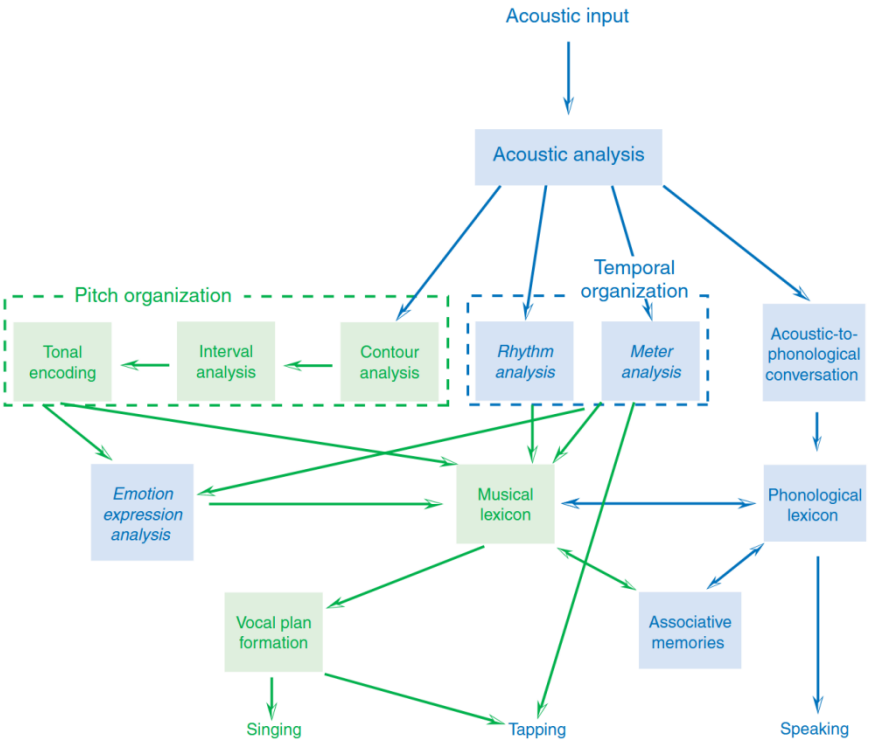


Figure 2.1 A modular model of music processing (adapted from Peretz & Coltheart, 2003).

Peretz's model fractionates music processing into a hierarchical system of distinct sub-components. As shown in Figure 1, each box represents a processing component, and arrows represent pathways of information flow or communication between processing components. At its core, the model proposes two parallel and largely independent pathways for processing the primary dimensions of music: pitch (including melodic contour and intervals) and rhythm (including meter and temporal grouping).

The primary evidence for this modular architecture is drawn from studies of individuals with amusia, a disorder of music processing. By employing the logic of double dissociation, research has shown that some individuals with brain damage can lose the

ability to process melodic information while retaining rhythm perception, and vice versa (Ayotte et al., 2002). Most critically for the broader music-language debate, a double dissociation has been established between musical and linguistic abilities. Numerous clinical cases have demonstrated that individuals with severe aphasia (impaired language) could have fully preserved musical functions, while those with amusia showed no corresponding deficits in language (Peretz, 2016). This evidence supported the modularity argument, suggesting that music is handled by a dedicated and encapsulated system.

This model also finds converging support from neuroimaging studies in healthy adults, which have revealed distinct neural activation patterns for different types of auditory stimuli. Seminal work from the Montreal Neurological Institute, for instance, found that while speech syllables activated higher-level secondary auditory cortices bilaterally, the perception of melodies preferentially engaged regions in the right hemisphere (Zatorre et al., 1992, 1994). This functional asymmetry, where the left auditory cortex specialized for the high temporal resolution typically required for speech and the right for the high spectral resolution of music, provides a clear neuroanatomical basis for two distinct processing systems (Zatorre et al., 2002).

Further reinforcing this view of specialized modules, research has identified a neural system dedicated not to a complex cognitive function, but to a specific class of auditory input: the human voice. In an fMRI study, Belin et al. (2000) identified regions in the STS, now known as the Temporal Voice Areas, that respond selectively to human vocal sounds over other complex sounds, including instrumental music. From a modularity perspective, the existence of the TVA provides compelling evidence for a domain-specific module tailored for one of the most biologically and socially significant signals for humans.

In sum, the modularity account is supported by a confluence of evidence from neuropsychology, neuroimaging, and voice perception research. This architecture, with its distinct and functionally specialized systems for music, language, and voice, provides a clear but challenging framework for understanding song perception. It raises a critical question: how do these potentially independent systems interact when processing a single stimulus that contains features relevant to all of them? Is a song parsed and its components routed to separate modules, or does its hybrid nature force a different mode of processing altogether? This question sits at the heart of evaluating the limits and explanatory power of the modularity account and sets the stage for the alternative models discussed below.

2.2.2 The AST Hypothesis: Early Acoustic Feature-Based Separation

Challenging the domain-based view of modularity, the Asymmetric Sampling in Time (AST) hypothesis provides a mechanistic explanation for hemispheric specialization that is rooted in the temporal and spectral characteristics of auditory signals (Poeppel, 2003). Rather than positing innate, domain-specific modules for music and speech, AST argues that left and right auditory cortices have evolved to process acoustic information at different temporal resolutions: the left hemisphere utilizes short integration windows (20–40 ms), optimal for the rapid modulations crucial to speech, while the right hemisphere favours longer windows (150–250 ms), suited to the slower temporal changes characteristic of melody and prosody. This division is closely linked to the intrinsic oscillatory properties of each hemisphere, with gamma-band oscillations predominating in the left and theta-band in the right (Giraud & Poeppel, 2012).

The AST hypothesis thus reframes lateralization as an emergent property of low-level auditory processing, rather than a simple reflection of cognitive domain boundaries.

Importantly, recent advances have provided empirical support for this model using more naturalistic stimuli. For example, Albouy et al. (2024) demonstrated that spectro-temporal modulation patterns, as an acoustic cue, can reliably distinguish song from speech at the earliest stages of auditory analysis, indicating that the functional separation between music and speech may originate from the brain's sensitivity to basic acoustic features, rather than from high-level cognitive specialization.

While the AST hypothesis has significantly advanced our understanding of how low-level auditory features are parsed and distributed across hemispheres, its explanatory scope is largely confined to the earliest stages of auditory analysis. AST accounts for how the brain initially separates speech and music based on their spectro-temporal characteristics, but it does not specify how they are further processed once they have entered higher-order cortical regions. In particular, the model leaves open questions about how syntax, whether linguistic or musical, is subsequently extracted and integrated from these early representations. To address these limitations, subsequent theoretical frameworks have focused on the neural mechanisms that support syntactic processing at more advanced stages.

2.2.3 Shared Resources and the SSIRH: Syntactic Integration Overlap

Moving beyond the early-stage, feature-based separation posited by AST, the Shared Syntactic Integration Resource Hypothesis (SSIRH) and related shared-resource accounts emphasized the overlap and interaction between music and speech at higher processing stages, particularly in the domain of structural integration. According to SSIRH (Patel, 2003), while music and speech maintain distinct representational systems (e.g., grammar, harmony), they recruit a common pool of neural resources for the processing of syntactic structure.

A range of evidence supports this view. Studies of congenital amusia indicate that deficits in pitch processing can affect not only music perception but also linguistic functions such as tone identification and prosody (Nan et al., 2010; Patel et al., 2005; Thompson et al., 2012). Event-related potential (ERP) research demonstrated that syntactic violations in both domains elicited similar neural responses, notably the P600 component, suggesting a domain-general structural integration mechanism (Koelsch, 2011; Patel, 2003). Rhythm processing provided further support: both music and speech relied on temporal sequencing, and neuroimaging studies have revealed overlapping activation in timing and motor networks (Tierney & Kraus, 2013).

Behavioural and training studies also indicate substantial cross-domain transfer. Musical training has been shown to enhance speech-in-noise perception, prosody processing, and phonological awareness, particularly in children and populations with atypical auditory development (Good et al., 2017; Kraus & Chandrasekaran, 2010). Patel's OPERA hypothesis (Patel, 2011) further specifies the conditions under which such transfer occurs, highlighting neural Overlap, Precision demands, Emotion, Repetition, and Attention. These findings collectively challenge the notion of strict modularity and suggest that, at least during syntactic integration, music and speech might share cognitive and neural resources.

2.2.4 The Three-Phase Model of Auditory Sentence Processing

A landmark in the neuroscience of language, Friederici's three-phase model (Friederici, 2002) provides a temporally and functionally precise blueprint for how the brain comprehends spoken sentences in real time, as shown in Figure 2. Crucially, this model decomposes sentence processing into three temporally distinct, neurally dissociable

stages, each supported by different brain regions and associated with specific ERP components.

Phase 1: Initial, Local Phrase Structure Building

Within about 150–200 ms, the brain rapidly assigns syntactic categories to each word and constructs basic local phrase structures. This first-pass syntactic parsing is primarily supported by the anterior Superior Temporal Gyrus (aSTG) in the left hemisphere. The ELAN (Early Left Anterior Negativity) ERP component indexes this stage, particularly in response to phrase structure violations.

Phase 2: Lexical-Semantic and Thematic Role Assignment

In the 300–500 ms window, often overlapping with the end of phase one, the brain retrieves word meanings from the mental lexicon and assigns thematic roles ("who did what to whom"). This process engages the middle and posterior temporal cortices, including Wernicke's area. The N400 ERP component reflects difficulties in semantic integration at this stage.

Phase 3: Final Integration and Syntactic Reanalysis

Between 600–900 ms, a late-stage integrative process occurs, supported by a network including the left Inferior Frontal Gyrus (Broca's area) and posterior Superior Temporal Gyrus. Here, the brain finalizes the syntactic structure, resolving ambiguities and repairing errors. The P600 ERP component is characteristic of this phase, especially during syntactic reanalysis or repair.

Friederici's model is particularly valuable in the current context because it suggests multi-stage processing for auditory perception, with precise neural and electrophysiological correlates. This allows for direct and systematic comparisons with models of music perception, such as Koelsch's (see the next section), and grounds the

multi-stage and hierarchical hypothesis in the temporal architecture of real-time speech comprehension.

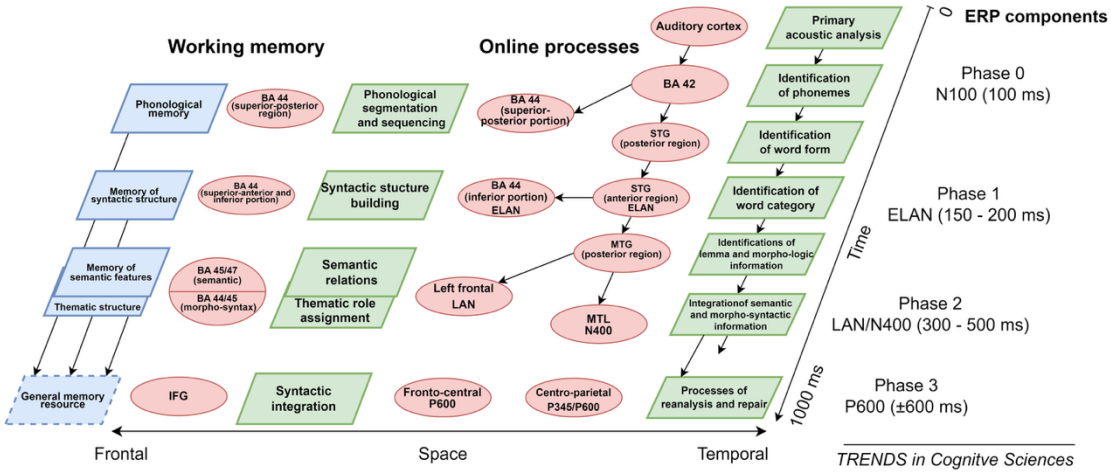


Figure 2.2 The Three-Phase Model of auditory sentence processing (adapted from Friederici, 2002, with changes in colour scheme and typography).

2.2.5 Koelsch's Comprehensive Model for Music Processing

Koelsch's (2011) comprehensive model offers an equally detailed multi-stage account of music perception, integrating insights from neuroimaging, electrophysiology, and cognitive psychology. This model, like Friederici's for speech, proposes that music processing unfolds in a series of hierarchically organized and partially overlapping stages, each associated with distinct brain networks, as shown in Figure 3.

Acoustic Analysis

Initial processing occurs in the brainstem, thalamus, and primary auditory cortex, focusing on basic acoustic features in a largely symmetric manner across hemispheres. Secondary auditory areas extract core musical features such as intervals, contours, and timbral qualities, further organizing the auditory input.

Structural (Syntactic) Processing

A fronto-temporal network, including the inferior frontal gyrus and superior temporal gyrus, is responsible for the processing of musical syntax, in which the rules governing the combination of musical elements. This stage closely parallels the syntactic integration stage in speech processing (as described by Friederici), and is associated with ERP components like the Early Right Anterior Negativity (ERAN) and P600, especially in response to musical irregularities.

Semantic and Emotional Processing

Posterior temporal regions, limbic structures, and the orbitofrontal cortex are engaged in the interpretation of musical meaning and the processing of emotional content.

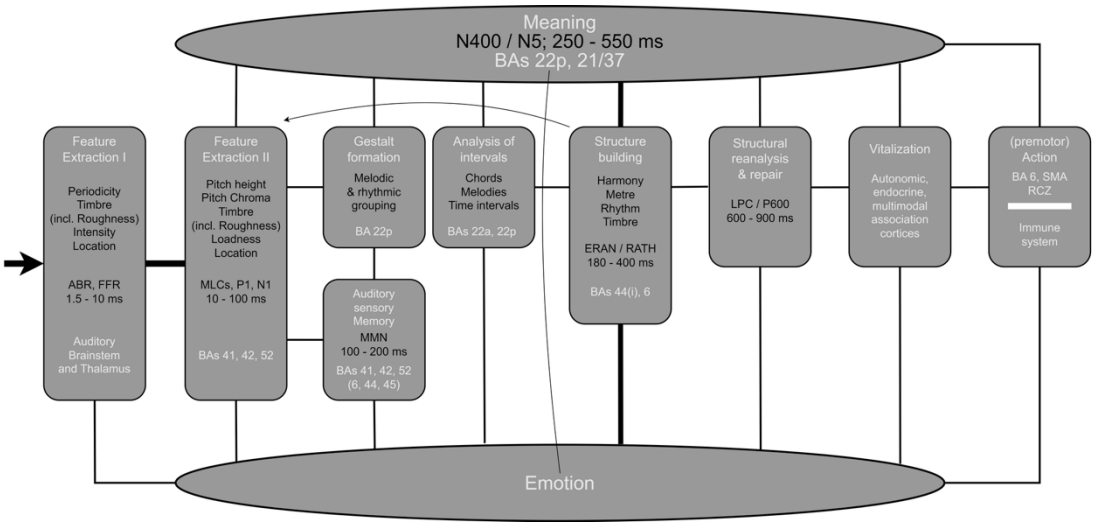


Figure 2.3 Comprehensive model for music processing (adapted from Koelsch, 2011, with changes in colour scheme and typography).

2.2.6 Model Comparison

Koelsch's model of music perception offers an integrative, multi-component framework that is both theoretically sophisticated and practically versatile. Crucially, this model is not only compatible with, but also capable of subsuming, earlier influential

accounts such as the AST and the SSIRH hypothesis. The AST hypothesis can be seen as specifying the foundational auditory processing principles at the model's earliest stages detailing how low-level spectro-temporal cues are parsed and transmitted to higher-order networks. SSIRH, in turn, aligns with Koelsch's conception of a fronto-temporal syntactic network, operating at intermediate levels of musical structure processing. In this way, Koelsch's model serves as a unifying framework that accommodates, as well as integrates, the insights of both AST and SSIRH within its multi-level architecture.

Importantly, Koelsch's model stands in parallel to Friederici's three-phase model of auditory sentence processing. Both frameworks delineate temporally organized, modular stages of neural processing, each supported by dedicated cortical networks and indexed by specific neural components (e.g., P600). They highlight the structural and functional analogies in the way the brain processes complex auditory signals, whether linguistic or musical. However, Koelsch's model is notably more comprehensive: while Friederici's model focuses primarily on the temporal and memory-related dynamics of sentence comprehension, Koelsch's account explicitly incorporates the processing of emotion, motivation, and reward, extending explanatory power to the affective dimensions of music that are not addressed in Friederici's linguistic-centred approach.

Despite these advances, both models share a key limitation: they have been developed largely in isolation from one another and are predominantly domain-specific, emphasizing either speech or music. They do not systematically address how the brain processes hybrid auditory signals such as song, in which musical and linguistic information are inseparably intertwined. This gap is particularly salient given the centrality of song to human culture and cognition (Mehr et al., 2019). The challenge is to develop unified, cross-domain models capable of capturing the interplay between

linguistic and musical information during the perception of song, which serves as a natural laboratory for investigating the mechanisms of multi-stage separation and integration in the auditory brain.

Table 2.1 A summary of five models explaining music and speech perception

Model	**Stage of Separation**	**Core Assumptions**	**Main Supporting Evidence**	**Limitations**
Modularity	System-level (late, domain-based)	Speech and music rely on distinct neural modules	Double dissociation, imaging, lateralization	Overlooks overlap/interactions; based on artificial stimuli
Voice Perception (TVA)	Early-to-intermediate (input-based, voice-specific)	The human brain has dedicated temporal voice areas for processing vocal sounds	fMRI showing selective STS activation for vocal vs. non-vocal sounds	Interaction with music pathways unclear
Asymmetric Sampling in Time (AST)	Early (acoustic feature-based)	Hemispheric specialization for temporal/spectral windows drives separation; left for speech (fast), right for music (slow)	fMRI/EEG; evidence for time/frequency windows, spectrotemporal separation	Integration at later stages not addressed

Shared Resources (SSIRH/OPERA)	Middle (syntactic/structural integration)	Overlapping neural resources for structural integration, domain-specific representations	ERP (P600) and behavioural evidence; transfer effects in training, overlap in rhythm/timing networks	Variability across stimuli and populations
Friederici's Three-Phase Language Model	Multi-stage, hierarchical (within speech)	Sentence processing unfolds in three phases: (1) early local syntax, (2) lexical-semantic, (3) reanalysis	ERP (ELAN/N400/P600), fMRI showing specific brain regions for each phase	Focuses on language; does not address music or hybrid processing; mainly tested on isolated speech
Koelsch's Comprehensive Model	Multi-stage, hierarchical	Integration of early acoustic, mid-level structural, and higher cognitive/emotional processing; dynamic network	Neuroimaging spanning multiple stages; aligns with AST and SSIRH elements	Routing of lyrics/melody unclear

As shown in the Table 1, these theoretical models locate the potential for music-language separation or integration at different levels of the processing hierarchy: AST at the level of early acoustic analysis, SSIRH at the stage of structural integration, Friederici's and Koelsch's models as attempts at multi-stage synthesis.

2.3 Song Perception: A Window into the Music-Speech Relationship

Given the theoretical landscape laid out above, song emerges as an ideal stimulus for testing multi-stage separation hypotheses. Basically, a song is a form of vocal music comprising melody, rhythm, and lyrics. It can be conceptualized as both sung speech and vocal music, with each term highlighting one aspect of its dual nature. This duality raises a fundamental theoretical question: is a song a simple mixture, where its components (lyrics and melody) are processed independently, or is it a compound, where these components interact to create a new, integrated perceptual entity greater than the sum of its parts (Jeffries et al., 2003)? This section will define song as a unique cognitive object, explore its position on a perceptual continuum with speech, review the evidence for the deep integration of its linguistic and musical components, and survey the current, unresolved debate regarding its neural processing.

More importantly, song perception invites us to consider a more nuanced, hierarchical model in which the boundaries between music and speech are dynamic. Thus, examining how song is processed offers direct insight into the mechanisms and stages at which music and speech might diverge or interact in the brain.

2.3.1 The Speech-Song Continuum

Traditional accounts have often conceptualized speech and song as discrete categories. However, contemporary research increasingly views them as endpoints on a continuous spectrum of human vocal communication (Vanden Bosch der Nederlanden et al., 2023).

This perceptual continuum is defined by gradients of acoustic features: prototypical speech typically exhibits variable pitch, gliding intonation, the absence of discrete intervals, unfixed durations, rapid tempo, and irregular rhythm. In contrast, prototypical song is characterized by stable, discrete pitches, intervallic structure, fixed durations, slower tempo, and metrically regular rhythm.

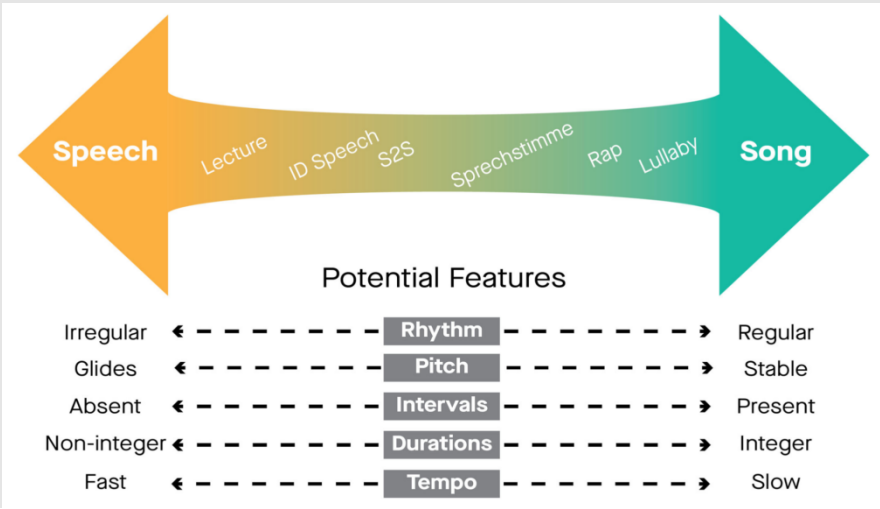


Figure 2.4 Continuum from speech to song. (Adopted from Vanden Bosch der Nederlanden et al., 2023).

Between the song and speech poles lie a range of intermediate vocalizations. Infant-directed speech (IDS), for example, uses exaggerated pitch, extended duration, and greater rhythmic regularity, making it more musical than adult-directed speech but still fundamentally linguistic in function (Cummins, 2013). Rap, or hip-hop music, which retain much of the prosody and intonation of speech, are closer to the speech endpoint, while chants and lullabies, with their repetitive melodies and simple structures, approach the song endpoint.

Compelling evidence for the continuity between speech and song comes from the speech-song illusion (Deutsch et al., 2011). In this phenomenon, listeners exposed to repeated loops of a spoken phrase begin to perceive it as being sung, rather than spoken.

Experimental data demonstrated that after ten presentations, participants were more likely to classify the stimulus as song, whereas a single presentation was perceived as speech. This finding underscored the fluidity of the perceptual boundary and highlights the role of acoustic cues such as repetition, pitch stability, and temporal regularity in shifting perception along the continuum.

Recent work by Albouy et al. (2024) further supported the view that music and speech were not strictly separate domains, but rather occupied different regions of a common spectro-temporal modulation space. Their findings suggested that song and speech were distinguished at the earliest stages of auditory processing by their specific acoustic structures, thus providing a mechanistic basis for the perceptual continuum.

2.3.2 The Deep Integration of Melody and Lyrics in Songs

The perception of song as an integrated "compound" rather than a simple "mixture" is most evident in the relationship between melody and lyrics, particularly in tonal languages where pitch height and pitch contour of a word could distinguish the semantic meanings. Consequently, a central question is how the melody of a song reflects and accommodates these linguistic components. The section demonstrates a complex but rule-governed integration that occurs at the levels of composition, performance, and perception.

Compositional Principles

The tone-tune association, or the principle that musical melody often mirrors lexical tone contours, has been studied for nearly a century (Herzog, 1934; Chao, 1956). Early qualitative observations gave way to rigorous quantitative studies, revealing that the degree of parallelism between word tones and melody is highly variable across languages and genres. For instance, while Mandarin pop songs and Beijing opera

frequently prioritize musical structure over tonal fidelity (Chao, 1956; Stock, 1999), Thai lullabies may show nearly perfect tone-tune mapping, in contrast to Thai pop songs, where mapping may drop to 59% (List, 1961). This suggests that tone-tune integration is a flexible, context- and genre-dependent process, not a universal rule.

The complexity is accentuated with contour tones. In languages like Cantonese and Vietnamese, research shows that the endpoint of a tone is most critical for determining the subsequent melodic transition (Chan, 1987; Kirby & Ladd, 2016; Wong & Diehl, 2002), whereas in Thai, both onset and offset may be important (Ho, 2006). Thus, the specific mechanisms of integration are speech dependent.

Performance Practice

Integration is not just theoretical or compositional; it is enacted in performance. Some composers explicitly encode linguistic-melodic integration in their scores. For example, Tan's *Trusting the Rainbows* (信任彩虹) requires performers to add a rising grace note for syllables with rising tones, making the preservation of speech contour an explicit part of musical execution.

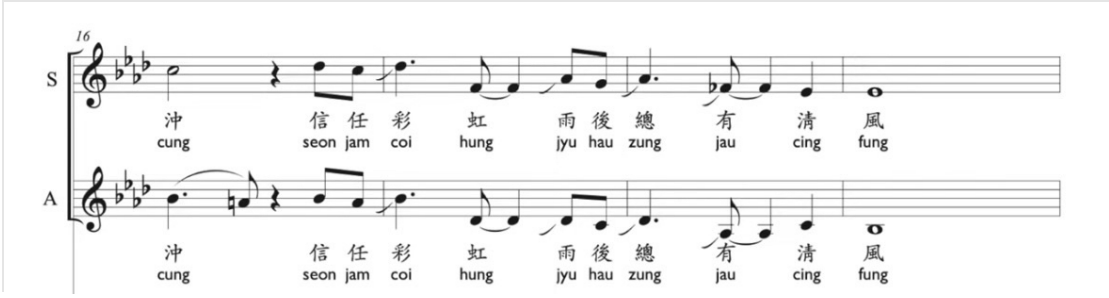


Figure 2.5 An excerpt of Tan's Trusting the Rainbows. The rising symbol before the note indicates that a semitone grace note should be sung, as in the syllable "coi", "jyu", "zung", and "jau".

Even without explicit notation, Cantonese singers often realize tone-melody integration automatically. M. Schellenberg & Gick (2018) found that Cantonese singers, when

performing a rising tone on a sustained note, would naturally produce a subtle upward pitch glide, embedding linguistic features within the musical structure.

Perceptual Consequences

This deep integration affects listeners at multiple levels. Wong & Diehl (2002) demonstrated that the melodic context preceding a syllable can bias the perception of its lexical tone, with high preceding melodies making a target more likely to be perceived as low, and vice versa. Furthermore, melody and lyrics are encoded in memory as integrated representations. Serafine (1984) found that familiarity judgments for lyrics are influenced by the melodies with which they are paired, suggesting that song is stored as a compound object rather than two independent objects. These findings collectively demonstrate that song represents a complex, rule-governed integration of linguistic and musical information at stages from composition, performance, to perception.

2.3.3 Song Perception: Familiarity, Memory, Emotion, and Neural Correlates

Beyond compositional and perceptual integration, higher-level cognitive factors such as familiarity, memory, and emotion play a profound role in song perception. Songs are not just processed as combinations of lyrics and melody; they are privileged in the mind and brain, often exhibiting properties that surpass those of speech or music alone.

Familiarity and Memory

A robust finding in the literature is the "song memory advantage." Songs are generally better remembered than spoken phrases or instrumental music, a phenomenon attributed to the binding of melody and verbal content (Racette & Peretz, 2007; Wallace, 1994). This advantage was observed in both short-term and long-term memory tasks and particularly pronounced when lyrics and melody were learned together rather than

separately (Serafine, 1984). The integrated encoding of lyrics and melody facilitates not only recall but also recognition and associative learning, making songs powerful tools for education and cultural transmission.

Emotion and Engagement

Emotion is another critical factor that distinguishes song from speech and instrumental music. Songs are more likely to evoke strong emotional responses, both subjectively and physiologically (Juslin & Laukka, 2003). The emotional salience of song is linked to its combined use of linguistic meaning, melodic contour, rhythm, and often personal or cultural associations. The emotional effect of music can enhance memory for songs (Belfi et al., 2016), suggesting a link between affective and mnemonic processes.

Neural Correlates

Neuroimaging studies have provided mixed but intriguing evidence regarding the neural processing of song. Studies on the general account suggest that song perception engages a widespread, overlapping network of brain regions. For instance, Schön et al. (2010) conducted two fMRI experiments among French listeners using 100 pairs of tri-syllabic words and 100 pairs of three-note melodies. Song stimuli were constructed by the pairing of words and melodies. Participants were instructed to make same-different judgements on song pairs including pairs having the same melody with different lyrics, having the same lyrics with different melodies, or having the same melody and lyrics. They discovered that the activated brain regions during the judgements were not confined to either a language network or a music network, supporting a domain-general processing account.

In contrast, the study by Albouy et al. (2020) used a sophisticated design to demonstrate a clear hemispheric double dissociation from fMRI results among French and English

listeners when speech and melody information were decoded separately, suggesting discrete processing of melody and speech. In their research, 10 sentences and 10 melodies were paired across each other to construct 100 song stimuli, whose spectral and temporal resolution were further degraded from 100% to 0%. Listeners should judge whether the song pairs sound the same or different in the perspective of sentence or melody. They found that decoding the linguistic content of a song relied on the left auditory cortex, while decoding the melodic content relied exclusively on the right auditory cortex, suggesting two discrete processing streams operate in parallel.

The design of Albouy et al. (2020) appeared to set a more difficult task for listeners than those, which might evoke different stages of auditory processing, compared with those of Schön et al. (2010). The inconsistent results may also stem from variations in stimulus designs (short phrases versus sentences; original recordings versus degraded samples; multiple singers versus one singer) and language backgrounds (English versus French).

More recently, a data-driven study identified a population of neurons that showed a supra-additive response specifically to song (Norman-Haignere et al., 2022). Intracranial responses to 165 auditory stimuli ranging from native and foreign speech, music, to other nonspeech sounds were analysed. The findings revealed a supra-additive response specifically for song processing accompanied by background music compared with synthesized songs or instrumental music, indicating the presence of a population of neurons exhibiting song selectivity. However, it remains unclear how *a capella* songs, which involve vocal music without instrumental accompaniment, would elicit similar neural responses. Exploring the perception of *a capella* songs would exclude the influence of instrumental music and provide insights into the continuity between music and speech based on the acoustic similarities between singing and spoken voice.

The divergent findings of previous song studies may stem from differences in stimuli (*a cappella* vs. accompanied, short phrases vs. sentences), tasks (passive listening vs. active judgment), and analytical methods. While these methodological differences offer plausible explanations for such divergent results, they highlight a more fundamental issue: the body of research on the cognitive neuroscience of song perception is still limited. With a small number of studies using varied paradigms, it is difficult to attribute inconsistencies to specific factors definitively. Thus, a comprehensive neural model for song perception has yet to be established, indicating a clear need for further systematic investigation.

As reviewed above, song perception is shaped not only by early acoustic but also by higher-order cognitive and affective factors, including familiarity, memory, and emotion. These dimensions collectively underscore the privileged status of song in human cognition, as well as the complex and multi-layered nature of processing.

In summary, the existing literature has identified multiple potential processing nodes where music and speech may diverge, from early acoustic analysis, through mid-level syntactic and semantic integration, to late-stage modulation by familiarity, memory, and emotion. However, no single study has yet systematically investigated these separation mechanisms within a unified framework that spans all processing levels. The present research is designed to fill this gap, using song as a natural, ecologically valid tool to probe the multi-stage architecture of music and speech processing in the brain.

3 Categorical Perception of Song-to-speech Continuum

3.1 Introduction

Human vocalizations occupy a complex acoustic continuum, ranging from prototypical speech through intermediate forms such as infant-directed speech (IDS), chant, and rap, and ultimately to prototypical song. This perceptual continuum is defined by gradients of acoustic features. Prototypical speech is typically characterized by a gliding f0 contour, the absence of discrete pitch intervals, variable syllable durations, and irregular rhythm. In contrast, prototypical song is characterized by stable, discrete pitches that form a clear intervallic structure, isochronous or metrically regular rhythm, and often slower tempo (Fitch, 2006; Vanden Bosch der Nederlanden et al., 2023). The perceptual fluidity between these endpoints could be demonstrated by the speech-song illusion, where the simple repetition of a spoken phrase can be perceived as sung, indicating that the perceptual boundary is context-dependent and not acoustically fixed (Deutsch et al., 2011).

Despite this physical continuity, listeners across cultures could effortlessly categorize vocalizations as speech or song. Developmental research has shown that this categorical skill is well-established by late childhood, with eight-year-old children's performance being comparable to that of adults (Vanden Bosch der Nederlanden et al., 2023). This perceptual duality, as shown by a discrete categorical experience arising from continuous acoustic input, presents a significant question of how the human brain impose perceptual boundaries upon an acoustic continuum.

Recent work has begun to reveal the neural and acoustic correlates underlying this ability of distinction. For example, Albouy et al. (2020) demonstrated that speech and song occupy opposite ends of a spectro-temporal modulation space and that these

categories are reliably decoded in the early auditory cortices. However, their approach relied on manipulating the spectral or temporal complexity of song stimuli rather than presenting a true acoustic continuum between song and speech. As such, while these findings support the existence of robust acoustic and neural markers, they did not directly address how listeners perceive sounds that fall along the continuum itself. Whether the perceptual boundary is sharp and categorical or gradual and continuous remains an open and critical question.

To address this gap, the present study adopts the Categorical Perception (CP) paradigm, a classic psychophysical approach for determining whether perception across a continuum is discrete or continuous. Categorical perception is evidenced by a sharp sigmoidal identification function between categories and a corresponding peak in discrimination accuracy at this boundary; this pattern indicates that perception is wrapped by the existence of categories, rather than tracking acoustic distance alone (Repp, 1984).

CP has been robustly documented in both speech (e.g., the perception of consonants, vowels, and lexical tones; Liberman et al., 1957) and music (e.g., pitch, intervals, and rhythm; Siegel & Siegel, 1977). However, comparative studies are often confounded by differences in sound source, particularly because human vocalizations and instrumental music differ fundamentally in both timbre and mechanisms involved in production.

To provide a direct and ecologically valid comparison, this study focuses exclusively on human vocalizations, thereby holding the production source constant. This approach is informed by the vocal source hypothesis, which posits that the brain possesses specialized neural mechanisms for processing the human voice (Belin et al., 2000). By

using only vocal stimuli, we can construct a true speech-song continuum unconfounded by non-vocal timbral differences.

The context of Cantonese is especially valuable for this purpose. The intrinsic and rule-governed integration of linguistic pitch (lexical tone) and musical pitch (melody) in Cantonese song is well-documented at the compositional, performance, and perceptual levels (Wong & Diehl, 2002). This property enables the precise, ecologically valid synthesis of stimuli spanning the speech-song continuum.

The central research question is how listeners perceive a continuous acoustic transition from speech to song. Two mutually exclusive hypotheses are tested. The categorical hypothesis predicts that listeners will exhibit classic CP effects, where a sharp categorical boundary in identification and a discrimination accuracy peak will be observed at this boundary. Such a finding would indicate that speech and song are represented as discrete perceptual categories, providing support for models that posit functionally distinct representations. In contrast, the non-categorical hypothesis predicts a gradual identification slope and/or discrimination performance that reflects acoustic distance rather than the relationship between categories. This outcome would support a more dynamic, continuum-based model of auditory cognition where category labels are emergent properties of acoustic analysis.

Beyond this primary question, the present study also examines top-down factors that might modulate perceptual categorization. Song familiarity is known to enhance memory, a phenomenon attributed to the integrated encoding of verbal and melodic information which creates a richer, more robust memory trace (Wallace, 1994; Serafine et al., 1984). Likewise, individual differences in musicality, including both professional training, objective musical competence, and musical sophistication, could shape the precision of auditory perception. Notably, musical training has been shown to enhance

categorical perception of speech sounds, suggesting a cross-domain transfer of perceptual skills (Bidelman et al., 2013). However, musical training has been considered only one aspect of a person's musicality. As indicated by previous research, there might be individuals who are gifted but musically untrained (Law & Zentner, 2012). To account for its heterogeneity, quantification of individual's musicality will be comprehensively measured using self-report data on musical training history as well as standardized tools such as the Profile of Music Perception Skills (PROMS, Zentner & Strauss, 2017) and the Goldsmiths Musical Sophistication Index (Gold-MSI, Lin et al., 2021), enabling detailed group comparisons.

In summary, this study directly tests whether the perceptual transition between speech and song is categorical in the CP frame and explores the modulatory roles of familiarity and individual's musicality. By leveraging the CP paradigm and carefully controlled vocal stimuli, it aims to clarify the mechanisms by which the brain resolves one of the most fundamental boundaries in human auditory experience.

3.2 Methods

3.2.1 Participants

Sixty-six Cantonese native speakers (37 females, 20.71 ± 1.82 yrs) have been recruited in the present study. All participants were recruited from the Hong Kong Polytechnic University. None of them majored in linguistics or psychology. They are all right-handed assessed by the Edinburgh handedness inventory (Oldfield, 1971). None of them reported any hearing impairment, speech or language-related disabilities, or brain injuries. Among them, thirty-seven were labelled as musical training group, as they self-reported to have received at least six years of formal musical training, including piano, violin, flute, vocal, or drums. Most members of this group have received more

than one type of instrument. Participants who reported no musical training experience were categorized as the control group. Table 2 shows the comparison between these two groups. The study received approval from the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. All participants have read and signed an informed consent form before the experiment began. They were given financial compensation after the experiment.

Table 3.1 Detailed information of participants of Experiment 1

	Music training group	Control group
Number	37 (23 females)	29 (14 females)
Age (SD)	20.65 (1.9)	20.71 (1.82)
Onset of training (SD)	6.59 (2.52)	N.A.
Lengths of training (SD)	10.16 (3.38)	N.A.

Power analysis

Prior to data collection, we conducted an a priori power analysis to determine the optimal sample size required to detect anticipated effects. Using the *pwr* package in R, we calculated required sample sizes across different effect sizes and analysis types, targeting a conventional power level of 0.80 with a significance level of 0.05.

The power analysis revealed substantial variation in required sample sizes (see Appendix Table 1). For medium-sized main effects (e.g., the influence of musical ability on boundary position), approximately 55 participants would be needed, whereas detecting small interaction effects (e.g., the interaction between musical ability and familiarity) would require up to 550 participants. The median required sample size across all analyses was 54 participants.

Given resource constraints and research priorities, we targeted a sample of 66 participants, which provides adequate power for most main effect analyses. With this sample size, analyses of medium effect sizes ($f^2 = 0.15$) for main effects (such as the effect of PROMS on boundary position) achieve a power of 0.873; for large effects ($f^2 = 0.35$), power reaches 0.997. However, for small interaction effects ($f^2 = 0.02$), the power is limited to 0.133, indicating that the study may lack sufficient sensitivity to detect subtle interaction effects. For medium-sized effects ($d = 0.5$) in the discrimination task, power was calculated at 0.516.

Based on these analyses, we determined that our sample size was sufficient to reliably address the primary research questions regarding the perceptual pattern of speech-song continuum and the influence of musical ability on speech-song boundary perception, while acknowledging that exploratory analyses of interaction effects would be interpreted with caution.

3.2.2 Stimuli

The singing materials were selected from eleven contemporary Cantonese songs, several of which are highly popular among Cantonese speakers, as listed in Table 3. Of these, three phrases consisted of four syllables, while the remaining eight were tri-syllabic. The excerpts of the selected songs along with their musical scores and lyrics were shown in the Appendix for reference. The lyrics from these songs were extracted to generate the corresponding spoken materials.

A female Cantonese singer with over ten years of formal vocal training was recruited to serve as both the vocalist and speaker for the experiment. She was instructed to deliver the text in a range of vocal registers. For each song, the musical key was selected to closely match the overall f0 of the spoken rendition. The singer performed the pieces

in a pop style at a tempo of 70 beats per minute, guided by a metronome. Each phrase was subsequently normalized to a duration of 1500 ms, and the intensity was adjusted to 70 dB SPL. By carefully controlling overall pitch height, duration, and singing style, the acoustic ambiguity between the song and speech materials was maximized, as shown in Table 4. The recording procedures were completed in a sound-proof booth using Audio-Technica AT2035 microphone connected to the Steinberg UR22mkII interface.

Table 3.2 Familiarity scores and acoustic differences of speech-to-song continua.

Song ID	Lyrics	HNR difference (dB)	Mean f0 difference (Hz)	Familiarity (SD)
0232	零二三二	4.264	15.463	3.86 (1.46)
AZY	愛自由	7.747	11.347	3.45 (1.75)
CSD	財神到	3.78	68.346	4.95 (0.21)
GHSY	光輝歲月	0.432	25.296	3.64 (1.55)
HXS	歡笑聲	7.43	40.505	2.82 (1.34)
NGW	你共我	2.041	3.167	4.27 (0.64)
QQQG	千千闕歌	1.313	24.927	3.05 (1.31)
TZD	天註定	9.992	37.721	2.59 (1.08)
XFX	新發現	0.514	55.627	1.95 (0.96)
XSD	新世代	1.472	40.294	1.50 (0.67)
YCH	迎春花	0.222	23.075	3.68 (1.31)

Table 3.3 The overall acoustic differences between song and speech materials.

	Song (SD)	Speech (SD)
Average pitch (Hz)	295.78 (56.14)	264.34 (51.55)
Harmonicity (dB)	21.85 (2.39)	18.55 (2.16)
Centre of Gravity (Hz)	497.77 (153.68)	737.36 (185.16)

The speech-song continua were constructed using TANDEM-STRAIGHT algorism (Kawahara & Morise, 2011), which enables to manipulate multiple acoustic features at the same time when constructing a continuum. Each continuum contains seven stimuli, with speech as the first step and song as the seventh step. Thus, the eleven sets of seven-step continua from speech to song were generated.

3.2.3 Task design and procedures

This study consists of three tasks: the familiarity test to song materials, the identification task, and the discrimination task. Additionally, participants were instructed to complete PROMS test and the Gold-MSI questionnaire after finishing these three tasks.

Familiarity task

In the familiarity task, participants were instructed to rate the familiarity scores towards each piece of song using a five-point scale, where five indicates they were most familiar, and one indicates least familiar with the song materials. Each song stimulus was only repeated twice with different key. And the familiarity test for each participant was conducted at least 10 days before the identification and discrimination task. The small number of repetitions of each song material and relatively long gap between this

familiarity test and the other two tasks were set to reduce possible effects evoked by this familiarity test.

Identification task

In the identification task, participants were instructed to identify whether they had heard was speech ("someone is talking") or music ("someone is singing") using a mouse. There were five repetitions for each stimulus, resulting in 11 continua × 7 steps × 5 repetitions = 385 trials. To avoid tiredness, participants were allowed to have a short break every 50 responses. Stimuli in the eleven continua were randomly presented, and behavioural responses were recorded using an open free Praat script shared by Peng (2023).

Discrimination task

In the discrimination task, participants were instructed to discriminate pairs of stimuli by clicking the "Same" or "Different" buttons using a mouse (AX design). The interstimulus interval (ISI) was 500 ms. For each speech-song continuum, 19 pairs of stimuli were included: 7 stimuli paired with themselves (i.e., #1-1, #2-2, #3-3, ... #6-6, #7-7) and 12 pairs of the neighbouring stimuli in forward order (i.e., #1-2, #2-3, ... #6-7) and reverse order (i.e., #2-1, #3-2, ... #7-6). To reduce the experimental time, five of the eleven continua were chosen as stimuli in this task, i.e., CSD, AZY, TZD, NGW, and XSD. There were five repetitions for each stimulus, resulting in 5 continua × 19× pairs × 5 repetitions = 475 trials. Similar as identification task, participants were allowed to have a short break every 50 responses to avoid tiredness. All pairs of stimuli were presented randomly.

Profile of Music Perception Skills (PROMS-S)

The current study aims to examine whether speech and music contexts would trigger different responses among groups with different musical backgrounds. We have collected the details of musical training information of each participant, including the onset and length of training as well as the instrument type, partially shown in Table 2. Apart from participants' self-report professional musical training details, we have also chosen the short version of Profile of Music Perception Skill (PROMS-S; Zentner & Strauss, 2017) as a battery to assess the participants' perceptual musical competence objectively. Compared with other batteries such as Montreal Battery Evaluation of Amusia (Peretz et al., 2003), the PROM-S could not only evaluate the musical ability of listeners among a wider range, but also assess multiple aspects of musical abilities including timbre, tuning, and embedded rhythm perception that have not been considered by most musical tests. The PROM-S contains eight subtests, examining different aspects relevant to musical properties: melody, rhythm, embedded rhythm, timbre, tuning, accent, tempo, and pitch. Within each subtest, there were 8 to 12 trials, resulting in around half an hour to be completed. After finishing the test, an overall score and the score of each subtest were obtained.

The Goldsmiths Musical Sophistication Index (Gold-MSI)

Previous studies have suggested that musical competence might not be the only criterion to indicate one's musicality. Rather, the engagement and appreciation of music could play a significant role (Hallam, 2010; Mas-Herrero et al., 2013). To examine whether this music sophistication would affect the performance of normalizing lexical tones, the Chinese version of Goldsmiths Musical Sophistication Index (Gold-MSI; Lin et al., 2021) was used. It is a self-reported questionnaire containing 39 questions using

a seven-point Likert scale to evaluate the degree of musical engagement. The questions included “I spend a lot of my free time doing music-related activities”, “I often pick certain music to motivate or excite me”, and “I am able to talk about the emotions that a piece of music evokes for me”. The score of the Gold-MSI was calculated using the template provided on its official website (https://shiny.gold-msi.org/gmsi_toplevel/).

3.2.4 Scoring and data analysis

For the identification task, the boundary position of category and the slope were calculated using a logistic regression equation as below, where P_1 represents the percentage of responses, b_0 is the intercept and b_1 is the identification slope. Boundary position, defined as the crossover point of the identification curve, was calculated when P_1 is equal to 50%.

$$log_e(\frac{P_1}{1-P_1}) = b_0 + b_1\chi$$

To examine factors influencing the boundary position of the speech-song continuum, we constructed a linear mixed-effects model with boundary position as the dependent variable. The model included individuals’ musicality (i.e., musical training, PROMS scores, and Gold-MSI scores) as the primary predictor of interest, with familiarity scores of each stimulus as a categorical predictor. The interaction between musicality and familiarity was also included to test whether musical expertise differentially affects boundary perception for familiar versus unfamiliar stimuli. The model incorporated random intercepts for participants to account for individual variability. The *lmer()* function from the *lme4* package was used to fit the model, and significance of fixed effects was assessed using likelihood ratio tests (Bates et al., 2015).

A similar linear mixed-effects model was used to analysis boundary slope. The model included musicality and familiarity as fixed effects, their interaction, and random intercepts for participants.

For the discrimination task, the stimuli pairs were grouped in units containing the pairs having stimuli with themselves and with their neighbouring stimulus. For example, the A-B unit consists of #A-A, #B-B, #A-B, and #B-A. There were six units to be calculated in total, based on the formula proposed by Xu et al. (2006).

Participants' acoustic sensitivity indicated by their ability to differentiate between adjacent stimuli along the speech-song continuum were analysed as the dependent variable. A linear mixed-effects model was adopted to analyse the effects on discrimination accuracy. The model included musicality, pair, and familiarity as fixed effects, along with their two-way and three-way interactions. Random intercepts for participants and stimulus pairs were included to account for individual differences and stimulus-specific effects.

We conducted correlation analyses to examine the relationship between different dependent measures. For these analyses, we used the *cor.test()* function from the *stats* package. Visualization of all results was performed using the *ggplot2* package (Wickham, 2016).

Model assumptions were checked using diagnostic plots from the performance package (Lüdecke et al., 2021). In cases where assumptions were violated, appropriate transformations were applied to the dependent variables or robust estimation methods were employed using the *robustlmm* package (Koller, 2016).

For all key analyses, we report both unstandardized and standardized effect sizes, along with 95% confidence intervals, calculated using the *parameters* package (Lüdecke et al., 2021). Post-hoc comparisons, where necessary, were conducted using the *emmeans*

package (Lenth, 2021) with Bonferroni correction for multiple comparisons. All statistical analyses were conducted using R (version 4.1.2) with a significance level set at $\alpha = .05$.

3.3 Results

3.3.1 Identification task

To determine the predictors of perceptual categorization along the speech-song continuum, a series of linear mixed-effects models were constructed to predict two key parameters derived from the identification task: the boundary position and the boundary slope. The analysis aimed to assess the relative contributions of two sets of predictors: Musicality of participants (binary musical training status, standardized PROMS scores, and standardized Gold-MSI scores) and Familiarity (average familiarity score among all participants for each song, and participant's individual familiarity score for each song pieces).

The analytical strategy involved two complementary approaches. First, an information-theoretic approach using model selection based on the Akaike Information Criterion with a correction for small sample sizes (AICc) was employed to identify the most parsimonious model from a set of 17 candidate models for each dependent variable. Second, a hypothesis-testing approach using nested model comparisons (ANOVA) was conducted to assess the overall significance of the Musicality and Familiarity predictor sets. All models included a random intercept for Subject to account for non-independence of data from the same participant.

Predictors of the Boundary Position

A set of 17 candidate models was specified to predict the identification boundary. The full comparison of these models, ranked by AICc, is presented in Appendix Table 2.

The model selection procedure provided a clear result, identifying model_b2 as the best-fitting model, which included only the standardized PROMS score as a fixed predictor. This model carried a substantial Akaike weight (w = .58), indicating it was considerably more likely to be the best model in the set compared to the next best competitor (model_b6; ΔAICc = 2.08, w = .21).

A detailed summary of this best-fitting model revealed that the PROMS score was a significant negative predictor of the boundary position (b = -0.55, SE = 0.12, t = -4.49, p < .001). This indicates that as participants' musical competence enhanced, their boundary position shifted towards the speech end of the continuum. The fixed effect of this model (PROMS score) explained approximately 7% of the variance in the boundary location (marginal R^2 = .07), while the full model including the random intercept for participants explained approximately 22% of the variance (conditional R^2 = .22). A confirmatory ANOVA supported this result, showing that the Musicality predictor set significantly improved model fit over a null model, $\chi^2(3) = 18.72$, $p < .001$, whereas the Familiarity set did not, $\chi^2(2) = 2.58$, $p = .276$.

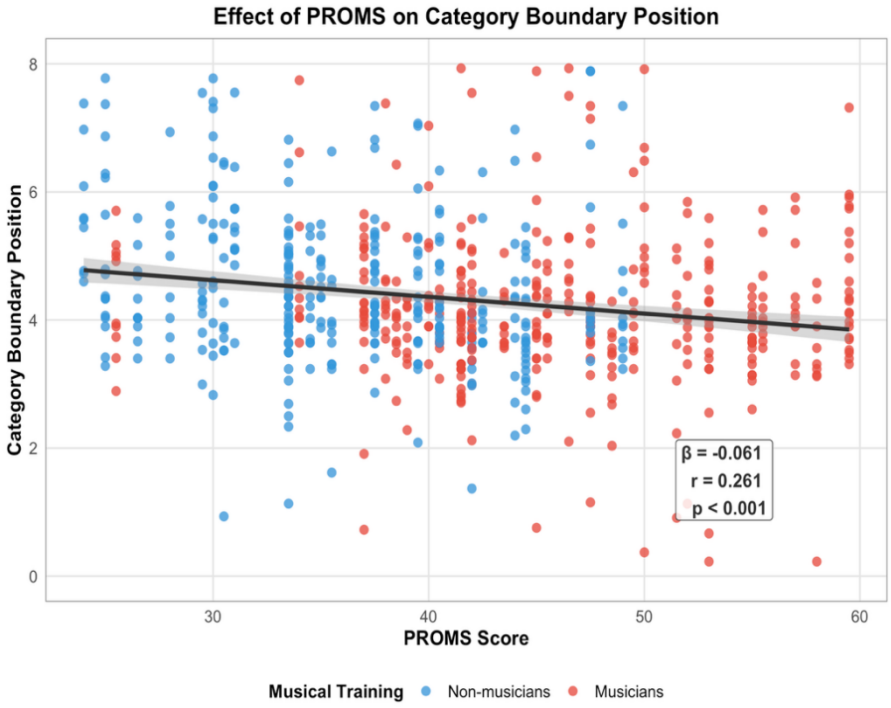


Figure 3.1 Relationship between PROMS scores and boundary position.

Note: Scatter plot shows individual data points coloured by musical training (blue = non-musicians, red = musicians). The black line represents the linear regression fit for all participants with 95% confidence interval shaded. Statistical values indicate the regression coefficient (β), correlation coefficient (r), and significance level (p). PROMS = Profile of Music Perception Skills. Higher boundary values indicate perception biased toward the song category.

Predictors of the Identification Slope

A parallel model selection procedure was conducted for the identification slope. The full comparison of the 17 candidate models is presented in Appendix Table 3. This analysis also identified the model with only the standardized PROMS score as the most parsimonious and best-fitting model (model_s2; w = .56). The next best model (model_s14; ΔAICc = 2.20, w = .19) offered substantially less support based on AICc values.

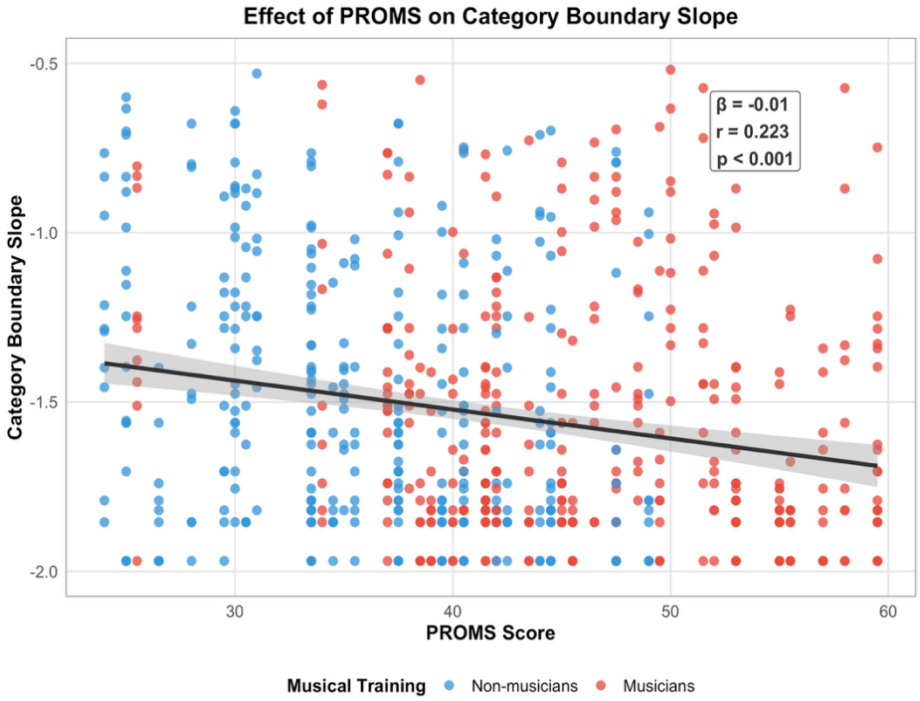


Figure 3.2 Relationship between PROMS scores and boundary slope

The summary of the best model showed that the standardized PROMS score was a significant negative predictor of the slope (b = -0.09, SE = 0.02, t = -3.78, p < .001). This finding suggests that higher musical competence is associated with a shallower identification slope, indicating a more continuous, less categorical pattern of perception. The fixed effect of PROMS score explained approximately 5% of the variance in the slope (marginal R^2 = .05), while the full model explained 20% (conditional R^2 = .20). However, the confirmatory ANOVA revealed a more complex picture. Both the Musicality predictor set, $\chi^2(3) = 16.36$, p < .001, and the Familiarity predictor set, $\chi^2(2) = 6.59$, p = .037, significantly improved the model fit over the null model. This suggests that while PROMS score is the single strongest and most parsimonious predictor, familiarity may also contribute to the variance in the identification slope.

Table 3.4 Summary of Fixed Effects for the Best-Fitting Linear Mixed-Effects Models

Dependent Variable	Predictor	*b*	*SE*	*t*	*p*
Boundary	(Intercept)	4.56	0.12	37.5	< .001
	PROMS Score (*z*)	-0.55	0.12	-4.49	< .001
Slope	(Intercept)	-1.51	0.02	-62.86	< .001
	PROMS Score (*z*)	-0.09	0.02	-3.78	< .001

Note. This table displays the unstandardized regression coefficients (b), standard errors (SE), t-values, and p-values for the fixed effects in the best-fitting models selected via AICc.

3.3.2 Discrimination task

A linear mixed-effects analysis was conducted to examine the factors predicting discrimination accuracy (ACC) in the discrimination task. The goal was to build a model that best explained discrimination accuracy based on the stimulus properties (Pair_n), participants' Musicality (musical training, PROMS scores, and Gold-MSI scores), and Familiarity (indF_z), while accounting for random variability across participants (Subject) and stimulus sets (continuum).

To determine the optimal set of predictors, a hierarchical model comparison was conducted using likelihood ratio tests. We began with a base model containing the main effects and sequentially added interaction terms to assess whether the increase in model complexity was justified by a significant improvement in model fit. The comparison of this series of nested models is presented in Appendix Table 4.

The results of the model comparison indicated that progressively adding interaction terms improved the model's explanatory power. The most critical improvement

occurred when the full set of theoretically motivated interaction terms (Pair_n × training, Pair_n × familiarity, familiarity × PROMS, and familiarity × Gold-MSI) was added to the main effects model. This step resulted in a highly significant improvement in model fit ($\chi^2(3) = 29.06$, $p < .001$). Subsequent additions of further interaction terms did not significantly improve the model fit (all ps > .05). Based on this observation, the model including all main effects and the significant two-way interactions was selected as the final, optimal model for interpretation. This model corresponds to model 7 in the analysis pipeline.

The final selected model containing fixed effects collectively accounted for 10.0% of the variance in accuracy (marginal $R^2 = .100$), while the full model including random effects accounted for 64.9% of the variance (conditional $R^2 = .649$). A detailed summary of the model's fixed effects is presented in Table 6.

Table 3.5 Summary of the Final Linear Mixed-Effects Model Predicting Discrimination Accuracy

Predictor	*b*	*SE*	*df*	*t*	*p*
Fixed Effects					
Intercept	0.641	0.053	5.87	12.17	< .001
Pair Position (Pair_n)	0.011	0.002	1910	5.91	< .001
Musical Training (Trained vs. Untrained)	0.007	0.023	93.92	0.31	0.759
Familiarity (indF_z)	-0.014	0.006	1965	-2.16	0.031
PROMS Score (PROMS_z)	0.04	0.009	65.86	4.22	< .001
Gold-MSI Score (Gold.MSI_z)	-0.004	0.01	65.05	-0.4	0.694

Pair_n × Musical Training	0.005	0.003	1910	1.89	0.059
Pair_n × Familiarity	0.005	0.001	1910	3.75	< .001
Familiarity × PROMS Score	-0.01	0.003	1964	-3.37	< .001
Familiarity × Gold-MSI Score	0.003	0.003	1962	0.94	0.346
Random Effects	**Variance**	***SD***			
Subject (Intercept)	0.003	0.056			
Continuum (Intercept)	0.013	0.113			
Residual	0.01	0.101			

Note. b = unstandardized coefficient estimate. Pair_n, indF_z, PROMS_z, and Gold.MSI_z were standardized continuous variables. Musical Training was a dummy-coded factor. df were estimated using the Satterthwaite's method.

The analysis revealed several significant main effects. There was a strong main effect of Pair along the speech-song continuum (Pair_n), such that discrimination accuracy increased as stimuli became more song-like ($b = 0.011$, $p < .001$). The PROMS scores was also a significant positive predictor, indicating that participants with higher objective musical competence performed more accurately overall ($b = 0.040$, $p < .001$). A significant negative main effect was found for Familiarity (indF_z), suggesting that participant's higher familiarity with the stimulus was associated with slightly lower discrimination accuracy ($b = -0.014$, $p = .031$). The main effects for Musical Training ($p = .759$) and the Gold-MSI score ($p = .694$) were not significant.

Figure 3.3 Discrimination Accuracy by Pair

Critically, the model identified two significant interaction effects. First, there was a significant interaction between Pair and Familiarity (Pair_n × indF_z; b = 0.005, p < .001). This indicates that the negative effect of Familiarity on accuracy was most pronounced for stimuli at the speech end of the continuum and was significantly attenuated for more song-like stimuli. Second, a significant interaction was found between Familiarity and PROMS score (indF_z × PROMS_z; b = -0.010, p < .001). This suggests that the performance advantage conferred by high musical competence (PROMS) was greatest for unfamiliar stimuli and was significantly reduced for stimuli with which participants were highly familiar. A marginally significant interaction between Pair and Musical Training was also observed (p = .059).

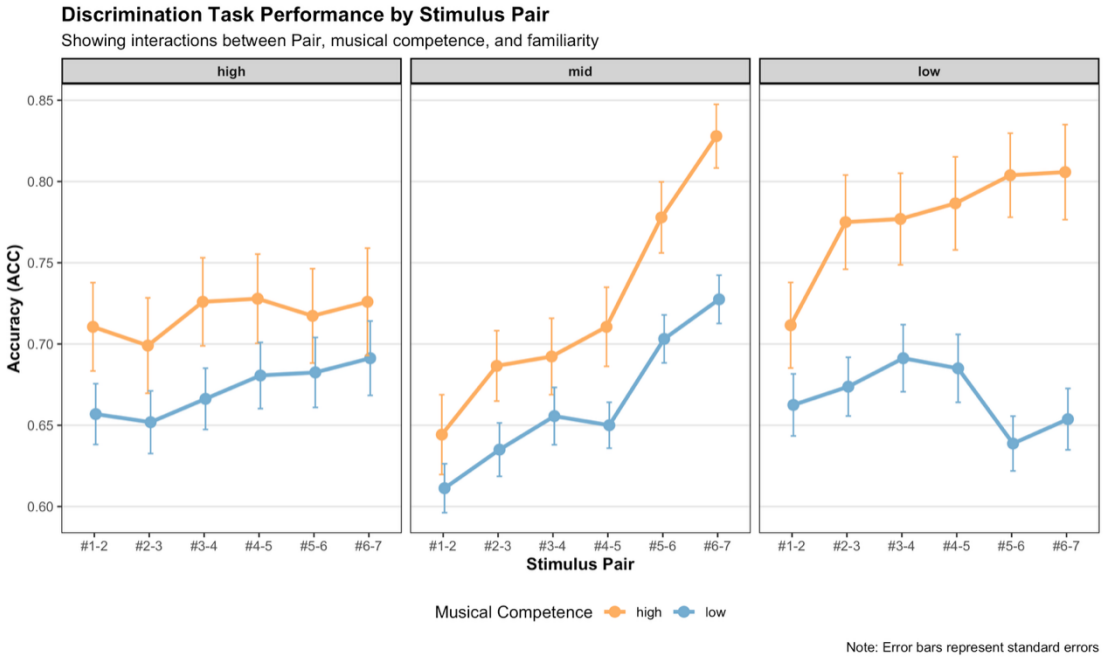


Figure 3.4 Discrimination Accuracy by Pair and Familiarity

3.4 Discussions

This study investigated the perceptual and cognitive mechanisms underlying the distinction between speech and song. By employing a Categorical Perception paradigm with carefully constructed speech-song continua, the research aimed to answer three questions: 1) Is the perceptual boundary between speech and song discrete and categorical, or continuous and graded? 2) What is the role of individual's musicality in shaping the perceptual pattern? 3) How does top-down knowledge, such as familiarity, modulate the processing of these hybrid vocalizations? The results provide a detailed picture that challenges simple dichotomous models and supports a more dynamic, continuum-based view of speech-song perception.

3.4.1 The absence of CP pattern

The primary research question concerned the nature of the perceptual boundary between Cantonese speech and song. The results from the identification and

discrimination task provided a clear answer: listeners' perception of the speech-song continuum was non-categorical. Unlike the sharp identification boundaries and corresponding discrimination peaks typically found in the perception of phonemes (Liberman et. al, 1957) or musical intervals (Siegel & Siegel, 1977), the transition from speech to song was perceived gradually. This conclusion is corroborated by a key pattern in the discrimination results. Rather than a discrimination peak at a category boundary, accuracy showed a continuous increase as stimuli progressed from speech-like to song-like, meaning that the acoustic sensitivity is increasing along the speech-song continuum.

This phenomenon can be explained by the fundamentally different cognitive goals associated with listening to speech versus music. The primary goal of speech perception is to achieve perceptual constancy, a process that requires the listener to ignore or suppress irrelevant acoustic variations (e.g., in a speaker's voice) in order to extract a stable, invariant linguistic message (Kuhl, 2004). In contrast, the aesthetic appreciation of music often requires listeners to attend to precisely these kinds of subtle variations in performance, such as a singer's timbre and expressive nuance (Gabrielsson, 2003). Therefore, the observed increase in discrimination accuracy for song-like stimuli seems to reflect a gradual rather than abrupt shift in listening strategy: from a speech-oriented mode that suppresses acoustic variance to a music-oriented mode that attends to it. The convergent evidence from the continuously increasing discrimination accuracy strongly supports the non-categorical hypothesis that the speech-song boundary is not a discrete perceptual line, consistent with phenomenological evidence like the speech-song illusion (Deutsch et al., 2011).

3.4.2 The Role of Musicality

The study revealed that objective musical competence, as measured by the PROMS test, was the most powerful predictor of listeners' performance in both identification and discrimination task. Two key findings emerged. First, participants with higher PROMS scores exhibited smaller boundary position that shifted toward the speech end of the continuum compared with those having lower PROMS scores, suggesting their higher sensitivity to the emergence of song-like acoustic features.

Second, and more critically, higher PROMS scores were associated with a shallower identification slope, indicating a less categorical mode of perception. This counter-intuitive finding suggests that musical expertise does not simply sharpen the boundary between speech and song; rather, it appears to foster a more continuous and analytical mode of listening. This may be because extensive training enhances the resolution of auditory representations at both cortical and subcortical levels (Bidelman & Krishnan, 2009), allowing expert listeners to perceive fine-grained acoustic details more veridically and rely less on the cognitive shortcut of forcing ambiguous stimuli into discrete categories. Notably, this effect was specific to the objective skill measure (PROMS), as neither the musical training nor the musical sophistication index (Gold-MSI) was significant predictors on their own. This has important methodological implications, highlighting the value of using objective psychophysical tests to quantify musicality over relying on training or subjective assessments (Zentner & Strauss, 2017; Müllensiefen et al., 2014).

3.4.3 The Complex Influence of Familiarity

The results from the discrimination task revealed the complex, modulatory role of top-down knowledge. The model showed a small but significant negative main effect of

familiarity, which was qualified by two significant interactions: the negative effect of familiarity was most pronounced for speech-like stimuli, and the performance advantage of high musical competence (PROMS) was greatest for unfamiliar stimuli. Together, these findings can be interpreted as evidence for a familiarity-induced shift in processing strategy. For unfamiliar stimuli, listeners are forced into a bottom-up, analytic listening mode, where they must attend closely to the acoustic details to perform the discrimination task. In this mode, higher musical competence with enhanced perceptual skill is advantageous. However, when a stimulus is highly familiar, it likely triggers a top-down, holistic recognition mode, activating a rich, integrated memory trace that binds lyrics and melody (Serafine et al., 1984; Wallace, 1994). This shift to a memory-based process, likely engaging different neural networks such as the medial prefrontal cortex (Janata, 2009), may reduce attention to the fine-grained acoustic details required for the discrimination task, thus slightly impairing performance and narrowing the gap between experts and non-experts.

4 Contextual Effect of Song in Lexical Tone Normalization

4.1 Introduction

Extrinsic contextual normalization is a fundamental perceptual mechanism that allows listeners to achieve perceptual constancy in the face of acoustic variability, a process particularly crucial for the identification of lexical tones in speech (Ladefoged & Broadbent, 1957). This phenomenon, in which the perception of a target sound is systematically shifted by the pitch of its surrounding acoustic context, is well-established. However, the underlying nature of this mechanism remains debate, primarily centred on whether it is a domain-general auditory process or a speech-specific mechanism.

The central debate has been informed by two competing lines of evidence. The domain-general account posits that normalization is a low-level auditory process sensitive to basic acoustic features, such as the mean f0 of the context. This view is supported by findings showing that the effect could occur across different languages (Wong, 1998) and even for non-native listeners with limited knowledge of the target language's tonal system (Fox & Qi, 1990). The most direct evidence for this account came from a study by Huang & Holt (2009), who reported that a non-speech context composed of sine waves could elicit a normalization effect for Mandarin tones.

In contrast, the speech-specific account argues that normalization relies on higher-level, linguistic processing. This view is supported by a larger body of research that has failed to replicate the non-speech context effect reported in Huang & Holt (2009). Studies in both Mandarin and Cantonese have found that various non-speech contexts including triangle waves, reversed speech, and iterated rippled noise elicit null or significantly weaker effects compared to the speech context (Chen & Peng, 2016; Francis et al., 2006;

K. Zhang et al., 2018). Crucially, a study using instrumental music (piano notes) as a context also failed to induce normalization for Cantonese tones, further bolstering the argument for a speech-specific mechanism (Tao et al., 2021).

However, a critical evaluation of the literature reveals a significant confound that complicates this debate: the sound source. The conclusion that speech normalization is domain-specific has been largely built on comparing human speech with non-vocal contexts (e.g., sine waves, instrumental music). The observed null effects could therefore be attributed not to a fundamental difference between speech and music processing, but to a more basic distinction between the processing of vocal versus instrumental sounds. This is consistent with the vocal source hypothesis, which posits the existence of specialized neural mechanisms in the superior temporal sulcus (STS) for processing the human voice (Belin et al., 2000). This raises a critical, unanswered question: what would happen if the musical context was also produced by a human voice?

Furthermore, the debate has often been framed as a binary, "yes-or-no" question of specificity. Recent research suggests a more nuanced, cumulative cue model may be more appropriate. For instance, C. Zhang & Chen (2016) demonstrated that the degree of normalization effect in speech will increasing as more linguistic cues are added to the context from phonetic information to phonological and semantic information. This suggests that normalization is not a simple switch but an additive process that leverages multiple levels of information.

Table 4.1 Four context types designed in C. Zhang & Chen (2016).

Context type	Information	Details
Nonspeech	Auditory	Triangle waves

Reversed speech context	Auditory + phonetic	Time-reverse the meaningful speech contexts
Meaningless speech context	Auditory + phonetic + phonological	呢錯視幣__. 'This mistake sees money __.'
Meaningful speech context	Auditory + phonetic + phonological + semantic & syntactic	呢個字係__. 'This word is __.'

The present study is designed to address these two critical gaps. First, by using vocal music (song and solfège) as contexts, we aim to disentangle the effect of the vocal source from the effect of abstract musical structure. Second, by adopting the cumulative cue framework, we aim to move beyond a binary view and ask a more precise question: if normalization is an additive process, what are the relative contributions of different cues within a vocal music signal? To do this, we designed five context types that systematically vary in their acoustic and linguistic content: (1) Instrumental Music (cello, a non-vocal baseline), (2) Solfège (vocal source, phonological cues + musical structure), (3) Song (vocal source, phonological cues + musical structure + semantic content), (4) Meaningless Speech (vocal source, phonological cues), and (5) Meaningful Speech (vocal source, semantic content)

Context type	Vocal	Musical structure	Phonological	Meaning
Cello	-	+	-	-
Solfège	+	+	+	-
Song	+	+	+	+
Meaningless SP	+	-	+	-

Meaningful SP	+	-	+	+

Based on the theoretical framework and research gaps identified above, the present study tests a set of specific hypotheses regarding the strength of the contextual normalization effect elicited by these different contexts. Our predictions are grounded in the literature suggesting that while the human voice is a critical cue, the decisive factor for speech-specific normalization is the processing of linguistically relevant phonological information.

First, we predict a primary vocal source effect. We hypothesize that the song and solfège contexts, as bearing vocal features, will elicit a significantly stronger normalization effect than the non-vocal instrumental music context.

Second, we predict a distinction between sung and spoken language based on the distinction of timbre and the presence of speech prosody. While the song context contains the necessary vocal and phonological cues to elicit normalization, its rigid musical structure and lack of natural speech prosody are expected to modulate the effect. Therefore, we hypothesize that the effect elicited by the song context will be significant, but smaller in magnitude than that elicited by a typical speech context.

Third, and most critically, we hypothesize that the function of the vocal signal, not just its acoustic presence, determines its influence on normalization. The syllables in the song context are processed for their linguistic meaning. In contrast, the syllables in the solfège context (e.g., *do*, *re*, *mi*), while phonetically and phonologically realized, are primarily interpreted within a musical system to convey musical meaning (i.e., scale degrees) rather than linguistic meaning. Although many of these syllables have homophones in Cantonese (e.g., *do* corresponds to '哆' /do1/), their deeply learned association with the musical scale is expected to dominate their processing. Therefore,

we predict that the solfège context will behave like instrumental music, failing to elicit a significant normalization effect. Its effect size is expected to be significantly weaker than that of the song context.

This predicted pattern of results (Speech > Song > Solfège ≈ Instrumental Music) would suggest that the specialization for speech normalization emerges not simply at the level of detecting a voice, but more precisely at the level of processing phonological information that is interpreted as being part of the linguistic system.

4.2 Methods

4.2.1 Participants

A total of 68 native Cantonese speakers (38 females, mean age = 20.71 ± 1.8 yrs) took part in this experiment. All participants were students at the Hong Kong Polytechnic University, with none pursuing degrees in linguistics or psychology. Right-handedness was confirmed for all participants using the Edinburgh handedness inventory (Oldfield, 1971). The screening process ensured that no participant had any history of hearing impairment, speech disorders, or neurological damage. Based on their musical background, participants were divided into two groups: 38 individuals with professional musical training (at least six years of formal training on instruments such as piano, violin, flute, voice, or percussion, with most having experience with multiple instruments) formed the musical training group, while the remaining 30 participants with no formal musical training constituted the control group. Comparative details between these groups are presented in Table 8. This research was conducted with approval from the Human Subjects Ethics Sub-committee at The Hong Kong Polytechnic University. All participants provided written informed consent before participating and received monetary compensation upon completion.

Table 4.2 Detailed information of participants of Experiment 2

	Music training group	Control group
Number	38 (23 females)	30 (15 females)
Age (SD)	20.5 (1.93)	20.97 (1.61)
Onset of training (SD)	6.58 (2.49)	N.A.
Lengths of training (SD)	10.16 (3.33)	N.A.

4.2.2 Stimuli

Stimuli were spoken and sung by two female Cantonese native speakers. They are both trained singers and long-term members of the choir of the Hong Kong Polytechnic University, of which one is soprano (F1), the other is alto (F2). The recording procedures were completed in a sound-proof booth using Audio-Technica AT2035 microphone connected to the Steinberg UR22mkII interface.

The design of the speech stimuli used in the current experiment was similar as Tao et al. (2021), Wong & Diehl (2003), and C. Zhang & Chen (2016). The monosyllable *ji* with mid-level tone (Tone 3) referring to the Cantonese word "meaning" (意) was chosen as the target stimuli. The other two level tones were recorded as the filler target words: *ji1* with the high-level tone (Tone 1) refers to the word "doctor" (醫), and *ji6* with the low-level tone (Tone 6) refers to the word "two" (二).

Different contexts were set to precede the target words with a jittering silence of 300-500ms. For each talker, two types of speech contexts and two types of singing contexts were recorded. Speech contexts consist of meaningful speech context and meaningless speech context. Four semantically neutral sentences were recorded as the meaningful speech context: 呢個字係 (/li55 ko33 tsi22 hɐi22/, "This word is meaning"), 下一個

係 (/hɐ22 yat55 ko33 hɐi22/, “The nest is”), 請留心聽 (/tsʰiŋ25 lɐu21 sɐm55 tʰiŋ55/, “Please listen carefully to”), and 我而家讀 (/ŋo23 ji21 kɐ55 tuk2/, “Now I will read”). Only the first sentence was used as experimental context. The other three sentences were used as a filler to reduce the familiarity effects of repetition. Next, all syllables were changed to *ji* to construct the meaningless speech context while their corresponding tone categories remained unchanged as those in meaningful speech context. For example, the experimental nonsense speech context was 醫意二二 (/ji55 ji33 ji22 ji22/, “doctor meaning two two”).

For the music contexts, there were singing with Cantonese lyrics (song context, henceforth), singing in solfège (solfège context, henceforth), and melodies played by cello (instrument context, henceforth). The musical notes performed by singers or cellos were designed as the closest pitch height in the chromatic scales to the corresponding syllable in the meaningful speech context, following the design in Tao et al. (2021). The melodies were shown in Appendix. Specifically, in the song context, singers were instructed to sing in the lyrics which are the same as in the meaningful speech context. In the solfège context, however, the lyrics were removed and the movable-do solfège was applied. For example, the corresponding experimental solfège context has been designed as *mi-do-la-so* for F1 and *mi-si-so-so* for F2. The instrument context was generated using *Garageband (10.4.11)* to synthesize melodies with a “cello” timbre. The same manipulation was conducted for the filler stimuli in each context.

To minimize the acoustic differences that are irrelevant to the current study, during the recording, the talkers’ production in speech, song, and solfège context was controlled at a tempo of 80 beats per minute indicated by a metronome. Next, the durations of each sentence were normalized to 1200 ms and the intensities were adjusted to 70 dB SPL.

The same manipulation in f0 was applied to five contexts after they were recorded or generated. The overall pitch trajectory of each sentence or melody was raised by 3 semitones to construct the raised context and was lowered by 3 semitones to construct the lowered context, aiming to elicit the contrastive contextual effect. In total, there were three different pitch shift conditions: raised, unshifted, and lowered. In the raised condition, listeners were expected to perceive the mid-level tone *ji*3 as low-level tone *ji*6. In contrast, they were expected to identify the same *ji*3 as *ji*1, the high-level tone.

4.2.3 Task design and procedures

In the warm-up and practice session, participants need to produce the syllables of the high, mid, and low-level tone categories clearly and be able to recognize these three tones in unshifted conditions. This could help to familiarize the following task and more importantly, ensure that Tone 3 and Tone 6 have not been merged in their tonal inventory. For the formal session, there would be a preceding context and a target word in each trial. Participants were instructed to listen carefully to the sound and identify the target word by labelling them into one of the following three categories: the high-level tone *ji*1, the mid-level tone *ji*3, and the low-level tone *ji*6. Before the context was presented, there was a 500 ms fixation in the middle of the screen. After the target stimulus was presented, participants needed to press the designated keys on the keyboard to indicate which word they had heard. The procedure is illustrated in Figure 10. The five contexts were assigned into five different blocks which were presented among the participants using a 55 Latin square design to reduce the order effect. Each stimulus was repeated nine times. In total, there were 5 context types (meaningful speech, nonsense speech, song, solfège, and instrument context) × 3 pitch shift

conditions plus one filler (raised, unshifted, lowered, and filler) × 2 talkers (the soprano F1 and alto F2) × 9 repetitions = 270 trials.

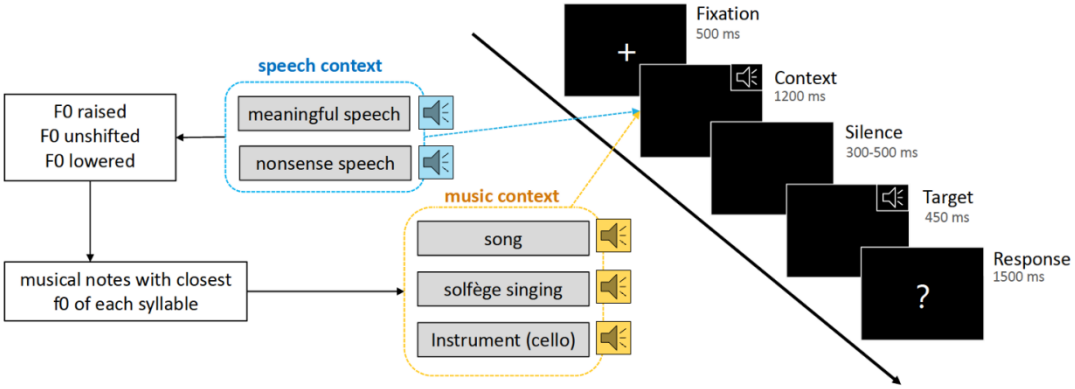


Figure 4.1 Experimental design of lexical tone normalization

4.2.4 Scoring and data analysis

For the word identification task, the expected identification rate (IR) for each context type was calculated following previous studies (Tao et al., 2021; Wong & Diehl, 2003). Specifically, we expected to observe a contrastive response opposite with the direction of pitch shift. For example, in meaningful speech context, the raised pitch shift condition was expected to elicit responses to identify the mid-level tone *ji*3 as the low-level tone *ji*6. The lowered pitch shift condition was expected to elicit more identification of high-level tone *ji*1. The higher the percentage of this response, the larger contextual effect could be considered.

4.3 Results

4.3.1 Power Analysis

We conducted a power analysis to evaluate the statistical sensitivity of our experimental design prior to analysing our results. This analysis was performed using the *pwr*

package in R (Champely, S., 2020), considering both our sample of 68 participants and the extensive repeated measures design (270 trials per participant).

The results demonstrated that our experimental design provided exceptional statistical power for detecting the effects of interest. For examining the main effect of context type (with 5 different contexts), our analysis indicated a power of 1.00 (100%) for detecting medium-sized effects ($f^2 = 0.15$). Notably, even for detecting small interaction effects ($f^2 = 0.05$) between context type and musical training, the calculated power remained at 1.00 (100%). This remarkably high statistical power stemmed primarily from the dense sampling approach of our design, which generated 18,360 total observations (68 participants × 270 trials).

Our power analysis confirmed that the current sample size, combined with the multiple observations per participant across various experimental conditions, offers substantially greater sensitivity than typically required for detecting both primary effects and interactions. This design ensured that even relatively subtle effects could be reliably detected in our statistical analyses. Given these power calculations, we determined that our experimental design was robustly equipped to address the critical questions regarding how different auditory contexts influence speech normalization and how musical training might modulate these effects.

A generalized linear model with a logit link function was employed to analyse the expected IR data. We conducted a comprehensive model selection process to determine the best predictors of response accuracy. The initial full model included Context type (meaningful speech, meaningless speech, song, solfège, and instrumental music context), musical Training, Competence, and Sophistication along with their interactions with Context type. Using a stepwise selection procedure based on AIC, implemented through the *stepAIC* function in R, we systematically evaluated various

model configurations to identify the most parsimonious model with the best fit. The final logistic regression model retained only Context type, Training, and their interaction.

4.3.2 Main Effect of Context Type

A significant main effect of Context type was observed in the logistic regression model, $\chi^2(4) = 1199$, $p < .001$. Performance varied significantly across the five different context types, with a clear hierarchical pattern of accuracy. Speech contexts (both meaningful and meaningless) yielded the highest accuracy, followed by song, with solfège and instrumental music (cello) showing the lowest accuracy. Pairwise comparisons revealed intermediate pattern of song context, which yielded significantly lower accuracy than both types of speech context (meaningful speech: $b = 0.24$, $SE = 0.07$, $z = 3.46$, $p = .005$; meaningless speech: $b = 0.37$, $SE = 0.07$, $z = 5.30$, $p < .001$), but significantly higher accuracy than both solfège ($b = 0.80$, $SE = 0.07$, $z = 11.44$, $p < .001$) and cello ($b = 0.78$, $SE = 0.07$, $z = 11.20$, $p < .001$).

The difference between meaningful speech and meaningless speech was not significant ($b = -0.13$, $SE = 0.07$, $z = -1.85$, $p = .347$), indicating comparable performance across both speech contexts. Similarly, no significant difference was observed between solfège and cello ($b = -0.02$, $SE = 0.07$, $z = -0.25$, $p = .999$), suggesting comparable performance in these two musical contexts. This pattern of results establishes a clear hierarchy of performance across context types: speech > song > solfège/instrumental music.

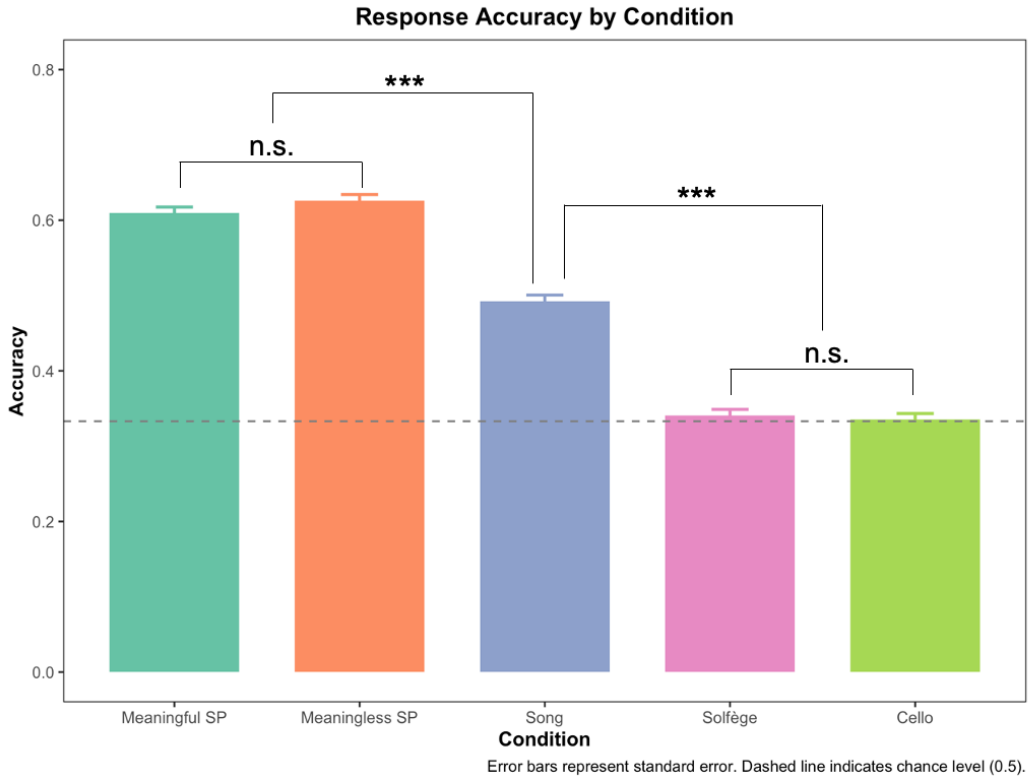


Figure 4.2 Identification rate in five conditions of normalization tasks

4.3.3 Effects of Musical Training

The logistic regression analysis revealed a significant main effect of Musical training ($b = 0.13$, $SE = 0.07$, $z = 1.98$, $p = .048$), indicating the context effect among participants with musical training was generally larger than those without musical training. However, this main effect should be interpreted in the context of the significant interaction with Context type ($\chi^2(4) = 25.49$, $p < .001$).

This interaction suggests that musical training affected the magnitude of differences between context types rather than changing the overall pattern of context effects. The interaction was strongest for song contexts ($b = -0.45$, SE = 0.09, $z = -4.75$, $p < .001$), indicating that the effect of musical training was substantially different in song contexts compared to meaningful speech. A smaller but still significant interaction was observed for cello contexts ($b = -0.20$, $SE = 0.10$, $z = -2.09$, $p = .037$), while the interactions for

solfège (b = -0.12, SE = 0.10, z = -1.27, p = .205) and meaningless speech (b = -0.11, SE = 0.10, z = -1.16, p = .248) were not statistically significant.

To understand how musical training modulated context effects, we compared the magnitude of differences between context types in each group. Musical training primarily affected the magnitude of differences between certain context types. The difference between speech prosody and song was substantially larger in musically trained participants (b = 0.69) compared to non-musically trained participants (b = 0.24). This 0.45 difference corresponds exactly to the significant interaction coefficient for Musical training × Song. Conversely, the difference between song and purely musical contexts was smaller in musically trained participants (Song vs. Solfège: b = 0.47; Song vs. Cello: b = 0.37) compared to non-musically trained participants (Song vs. Solfège: b = 0.80; Song vs. Cello: b = 0.78).

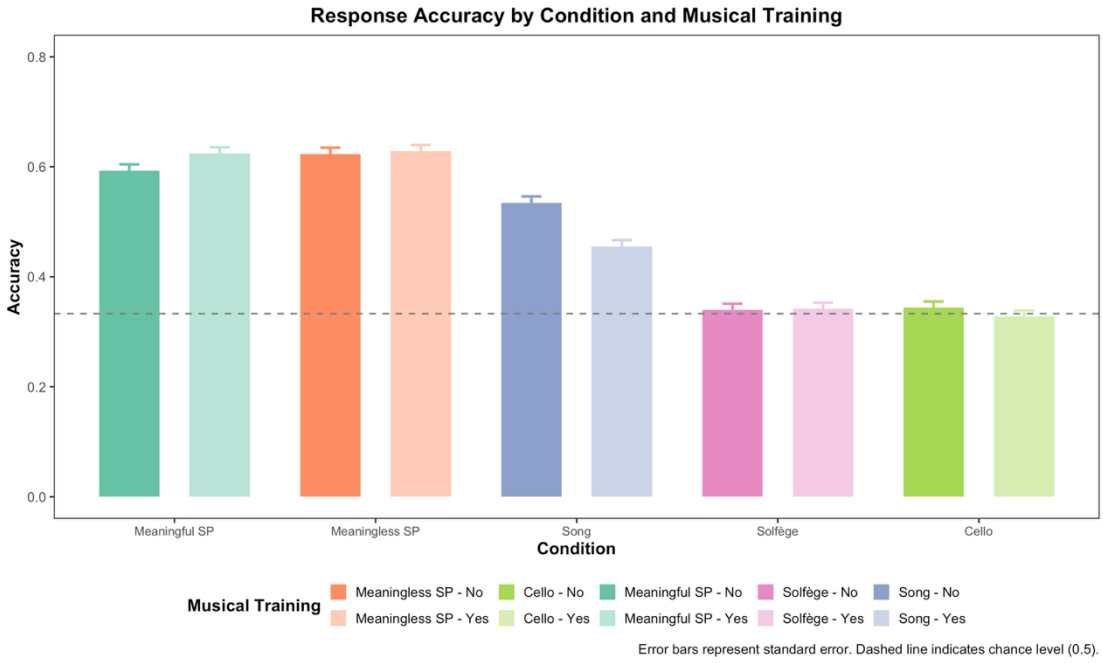


Figure 4.3 Identification rate by Condition and Training

Despite these differences in magnitude, the overall hierarchical pattern remained consistent across both groups: speech > song > solfège/cello. These findings indicate

that musical training did not fundamentally change the pattern of context effects across different context types but rather modulated the magnitude of differences between them. Specifically, musical training amplified the gap between speech and song contexts while reducing the gap between song and purely musical contexts.

Notably, the model selection process evaluated three different approaches to quantifying musical ability: Musical Training, standardize PROMS score, and the standardize MSI score. Despite the availability of these continuous measures, the stepwise selection process eliminated both PROMS and MSI scores from the final model, retaining only the training variable.

Table 4.3 The identification rate (%) in each condition and context type.

Context type	Raised (SD)	Unshifted (SD)	Lowered (SD)
Meaningful speech	43.6 (4)	66.8 (3)	73.3 (3)
Meaningless speech	47.2 (4)	78.7 (3)	59.6 (5)
Song	33.2 (4)	59 (3)	47.7 (4)
Solfège	27.3 (3)	47.4 (2)	28.1 (3)
Instrument	29.7 (2)	43.3 (2)	27.1 (3)

4.4 Discussions

This study investigated how different types of auditory contexts influence lexical tone normalization, focusing particularly on the relative contributions of vocal source, musical structure, and linguistic content to contextual processing. By systematically varying these elements across five context types, we aimed to disentangle their effects and test specific hypotheses about the mechanisms underlying speech normalization.

The results reveal a clear hierarchy of context effects and demonstrate that musical training modulates the magnitude of these effects in specific ways.

4.4.1 Hierarchy of Context Effects

The logistic regression analysis revealed a clear hierarchical pattern of context effects across the five context types: speech (both meaningful and meaningless) > song > solfège/instrumental music (cello). This pattern confirms our predictions and provides insights into the mechanisms of speech normalization.

Our first hypothesis predicted a primary vocal source effect, with song contexts eliciting stronger normalization effects than non-vocal instrumental music contexts. This prediction was supported by our findings, as song contexts consistently produced stronger effects than instrumental cello contexts. This aligns with previous research suggesting that the human voice serves as a critical cue for speech normalization (Levi et al., 2011; C. Zhang et al., 2016). However, the data reveal a more complex picture that goes beyond a simple vocal versus non-vocal distinction, as indicated by the data of solfège results.

Our second hypothesis predicted that while song contexts would elicit significant normalization effects, these effects would be weaker than those elicited by speech contexts due to the rigid musical structure and lack of natural speech prosody in songs. This prediction was also supported, as both types of speech context (meaningful and meaningless) produced significantly stronger context effects than song contexts. This finding suggests that natural speech prosody, with its linguistic-specific temporal and intonational patterns, plays a crucial role in facilitating speech normalization beyond what can be achieved by vocal cues alone.

Critically, our third hypothesis predicted that the function of the vocal signal, not just its acoustic presence, determines its influence on normalization. We expected that solfège contexts, despite containing vocal and phonetic elements, would fail to elicit strong normalization effects because the syllables in solfège are primarily interpreted within a musical system rather than a linguistic one. This prediction was strongly supported by our finding that solfège context behaved similarly to instrumental music context, with both eliciting significantly weaker effects than song contexts despite solfège containing vocal features. This pattern (Speech > Song > Solfège ≈ Instrumental Music) suggests that the specialization for speech normalization emerges not simply at the level of detecting a voice, but more precisely at the level of processing phonological information that is interpreted as being part of the linguistic system.

Additionally, no significant difference between meaningful and meaningless speech contexts was found in our study, suggesting that semantic content may play a less crucial role in speech normalization. This finding aligns with the theoretical framework suggesting that normalization primarily operates at the phonological level rather than the semantic level (Mitterer, 2006; Norris et al., 2003).

4.4.2 Revisit the Cumulative Cue Framework

Our results are best interpreted within the cumulative cue framework, which posits that speech normalization is not an all-or-nothing process but a graded one that leverages multiple levels of information. This perspective moves beyond a binary question of specificity. For instance, C. Zhang & Chen (2016) systematically demonstrated this additive principle by constructing a hierarchy of speech contexts (from reversed speech to meaningless and meaningful speech). They found that the normalization effect grew progressively stronger as more layers of linguistic information (phonetic, phonological,

semantic) were added. Our findings do not refute this model but rather support, extend, and refine it in two critical ways.

First, our results provide a clarification of what constitutes an additive linguistic cue. The critical dissociation between the song and solfège contexts is the key. This comparison cannot be explained by a simple lack of semantic meaning, as it is well-established that meaningless speech elicits robust, speech-like normalization (C. Zhang & Chen, 2016). The crucial difference, therefore, lies in the functional interpretation of the syllables. The syllables in a song, while set to music, are processed as components of the linguistic system to convey a message. In contrast, the syllables in the solfège context (e.g., *do, re, mi*) are primarily interpreted within a musical system to denote scale degrees. The failure to elicit a significant normalization effect of solfège contexts suggests a key refinement to the cumulative cue model: for a vocal-phonetic cue to be facilitatory and cumulative, it must be perceived as linguistically functional.

Second, our study demonstrates that the cumulative cue framework must account not only for facilitatory linguistic cues but also for competing or inhibitory non-linguistic cues. This is evidenced by the finding that the normalization effect for the song context was significantly weaker than for natural speech. While the song contains the necessary vocal and linguistic cues to trigger normalization, it also contains strong, non-speech-like musical structures (e.g., stable pitch, metrical rhythm). The intermediate effect elicited by song suggests a perceptual compromise, indicating that the presence of these salient musical cues actively modulates and attenuates the overall strength of the normalization effect.

In sum, our findings support a more dynamic view of the cumulative cue framework. The ultimate strength of the normalization effect appears to be the net outcome of a dynamic weighting of both facilitatory linguistic cues and competing musical-structural

cues present within a complex vocal signal, all filtered through the crucial lens of perceived linguistic function.

4.4.3 Musical Training and Context Processing

One of the most intriguing findings of this study is the significant interaction between musical training and context type. Musical training did not fundamentally alter the overall pattern of context effects but instead modulated the magnitude of differences between specific context types. Notably, musical training amplified the gap between speech prosody and song contexts while reducing the gap between song and purely musical contexts (solfège and cello).

This modulation effect suggests that musical training influences how individuals perceive and process song contexts. Rather than enhancing the linguistic aspects of songs, musical training appears to make individuals more sensitive to the musical elements within song contexts. For musically trained participants, song contexts seem to function more like musical stimuli than linguistic stimuli, as evidenced by the reduced gap between song and purely musical contextual effects. This interpretation aligns with the observation that the effect of song contexts was diminished in musically trained participants, making these contexts function more similarly to purely musical contexts (solfège and cello).

Our finding suggests that musical expertise may shift attention toward the musical structure and melodic information in songs, potentially at the expense of linguistic processing. Musically trained individuals may be more driven by the musical elements in song contexts, causing these contexts to be processed more like music and less like speech. This attentional bias toward musical features could explain why the

normalization effect of song contexts was weaker in musically trained participants compared to those without musical training.

Interestingly, our model selection process revealed that a simple binary classification of musical training status provided better predictive power than continuous measures of musical ability (PROMS and MSI). This suggests that professional musical training may lead to qualitative changes in auditory processing that are not fully captured by fine-grained measures of musical competence or sophistication index. These findings contribute to the growing literature on the effects of musical training on speech processing (Besson et al., 2007; Patel, 2011) by highlighting how musical expertise can redirect attention toward musical elements in complex auditory signals that contain both linguistic and musical information.

Our findings have several important theoretical implications. First, we support a domain-specific view of speech normalization that emphasizes the role of linguistic relevance rather than just acoustic similarity. The fact that solfège contexts, despite containing vocal elements, behaved similarly to instrumental music contexts suggests that the critical factor in normalization is not simply the presence of a human voice but the linguistic interpretation of that vocal signal. Second, our results highlight the importance of considering the function of different elements within a complex signal rather than just their physical properties. The same phonetic material (syllables) could have dramatically different effects on normalization depending on whether it is interpreted as part of the linguistic system (as in songs) or the musical system (as in solfège). Third, the interaction between musical training and context type suggests that expertise in music domain could modulate processing in speech domain, but in ways that are specific to the relationship between the two domains rather than as a general enhancement effect.

5 Neural Mechanisms of Song Processing and Familiarity Modulation

5.1 Introduction

The preceding behavioural investigations in this thesis have delineated the complex interface between speech and music. Experiment 1 (Acoustic-Perceptual) demonstrated that the boundary between speech and song is not categorical but continuous. Experiment 2 (Lexical Tone Normalization) revealed a hierarchical functional separation (Speech > Song > Solfège), suggesting that song occupies a hybrid ground where linguistic meaning constrains acoustic processing. These findings imply a multi-stage process: an initial, continuous analysis of acoustic features followed by a functional divergence driven by semantic content.

The critical remaining question is how the brain supports this hybrid processing and how higher-level cognitive factors, specifically prior knowledge (familiarity), modulate these mechanisms. This question is particularly pertinent in the context of tonal languages like Cantonese. In Cantonese songs, pitch serves a dual function: it is both the structural unit of musical melody and the distinguishing feature of lexical meaning. Uniquely, Cantonese songwriting follows relatively strict "tone-tune alignment" rules, creating a complex acoustic environment where lexical tones must parallel melodic pitch trends. This integration imposes a high cognitive demand on the brain to simultaneously decode musical structure and linguistic semantics.

To address this, the present study employs functional near-infrared spectroscopy (fNIRS) to investigate the neural correlates of song perception. We propose a Dual-Pathway Hypothesis modulated by familiarity. We posit that the brain does not process "familiarity" as a monolithic construct; rather, familiarity acts as a switch that shifts the brain's processing strategy based on the information richness of the signal:

The Skill Acquisition Pathway (Solfège): For vocalizations that possess musical structure but lack semantic meaning (Solfège), we hypothesize that processing relies on the dorsal auditory-motor stream for predictive simulation. Consequently, familiarity should induce Neural Efficiency, characterized by reduced activation as the pitch-syllable mapping becomes automatized.

The Active Retrieval Pathway (Song): For songs, where melody is inextricably bound to lyrics, we hypothesize that processing relies on the formation of a unified cognitive schema. Unlike solfège, familiarity with a song should trigger Active Retrieval and Integration. We predict this will be manifested as increased activation in prefrontal executive regions (specifically the Left Middle Frontal Gyrus, L-MFG) as the brain actively retrieves and integrates melodic, lyrical, and episodic information from long-term memory.

By investigating these hypotheses, this study aims to provide neural evidence for the multi-stage processing of song, clarifying how the brain integrates acoustic, linguistic, and higher-level cognitive information at the interface of music and speech.

5.2 Methods

5.2.1 Participants

Thirty native Cantonese speakers (16 females; mean age = 21.78 ± 1.59 yrs) were recruited for the fNIRS experiment. All participants were recruited from the Hong Kong Polytechnic University, and none were majoring in linguistics, psychology, or related fields. All were confirmed to be right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971). Participants reported no history of hearing impairment, speech or language-related disabilities, or neurological or psychiatric conditions.

A key criterion for this experiment was that all participants were non-musicians, established by self-report confirming no professional musical training. To objectively verify their non-expert status, their overall scores on the Profile of Music Perception Skills (PROMS) (M = 35.88, SD = 6.02) were compared to the scores of the musically trained group from Experiments 1 and 2 (M = 46.07, SD = 7.74). An independent samples t-test confirmed that the scores of the participants in the present experiment were significantly lower than those of the previously recruited trained musicians, $t(65) = 5.92$, $p < .001$.

This study received ethical approval from the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. All participants provided written informed consent before the experiment and received financial compensation upon its completion.

5.2.2 Materials

The auditory stimuli were designed to isolate the cognitive and neural contributions of melody, linguistic content, and familiarity in voice perception. Four distinct types of vocalizations were created: song, solfège, spoken lyrics, and spoken story. All stimuli were recorded by a single female native Cantonese speaker who is also a professionally trained singer.

Song Stimuli

Ten excerpts from contemporary Cantonese pop songs were selected to serve as the core musical materials (see Appendix). To investigate the role of familiarity, these were divided into two sets based on a priori popularity metrics. Five high-familiarity songs were selected from commercially successful albums and popular Cantonese music charts (e.g., based on online streaming data and public recognition). Five relatively lower-familiarity songs were selected from less commercially known artists or

operationally defined as having fewer than 10,000 views on major streaming platforms like YouTube. This selection process was designed to create a clear distinction in likely exposure for the participant population. The singer performed all ten excerpts *a cappella* to avoid confounds from instrumental accompaniment.

Solfège Stimuli

To isolate the melodic component while retaining the vocal source, a set of solfège stimuli was created. For each of the ten song excerpts, the singer replaced the original Cantonese lyrics with the corresponding solfège syllables (e.g., *do, re, mi*). Critically, the melody, rhythm, and overall vocal production style were kept identical to the original song performances. The purpose of this condition was to present a stimulus with a human voice and musical structure, but without lexical-semantic content, allowing for a direct comparison with the linguistic-meaningful song condition.

Spoken Lyrics Stimuli

To isolate the linguistic component, the same speaker/singer was recorded reading the lyrics from the ten song excerpts. She was instructed to speak in a natural, prosodically expressive manner, akin to reciting poetry, but without a fixed melodic contour or a metrically regular rhythm. The purpose of this condition was to present a stimulus containing the same vocal source and lexical-semantic content as the songs, but without the overarching musical structure.

Spoken Story Stimuli

To provide a baseline for ecologically valid, naturalistic speech, excerpts were taken from the Cantonese version of *The Little Prince* (*Le Petit Prince*). This text was chosen as it contains a rich mix of narration, dialogue, and monologue, representative of typical

spoken language. Several 20-second segments were extracted to serve as the natural speech condition.

All recordings were made using Audio-Technica AT2035 microphone connected to the Steinberg UR22mkII interface in a sound-attenuated booth at a sampling rate of 44.1 kHz and 16-bit resolution. To control for overall pitch differences, the musical key for each song performance was chosen to align with the average fundamental frequency (f0) range of the speaker's natural reading voice. While this provides a conservative level of control, it is acknowledged that the intrinsic acoustic differences between singing and speech, such as the wider pitch range and greater pitch stability in singing (Ozaki et al., 2024), were preserved to maintain the ecological validity of the stimuli. Each stimulus was normalized to 20 seconds. The root-mean-square (RMS) intensity of all stimuli was equalized to 70 dB SPL to ensure perceptual loudness was consistent across all conditions.

5.2.3 Procedures

Upon arrival at the laboratory, participants provided written informed consent and were seated in a comfortable chair in a sound-attenuated and dimly lit room. The fNIRS cap and system were then fitted to the participant's head. The probe set, consisting of a specific number of sources and detectors, was configured to cover bilateral temporal and inferior frontal regions. Auditory stimuli were delivered via Etymotic Research ER-1 Insert Earphones fitted with disposable foam eartips (ER3-14A), and participants' responses were collected via a keyboard.

The experimental session began with a 3-minute resting-state recording to establish a stable physiological baseline. Following this, the main experimental task commenced,

which was programmed and presented using PsychoPy software (Peirce et al., 2019). Before the first trial, an initial 30-second silent baseline period was recorded.

Each trial followed a consistent structure. Participants first listened to a 20-second auditory stimulus. Immediately following the stimulus, two sequential two-alternative forced-choice (2AFC) questions appeared on the screen. The first question asked, "Have you heard this specific content before?" (Yes/No), and the second asked, "Do you feel familiar with this content?" (Familiar/Unfamiliar). Participants were instructed to respond as accurately as possible by pressing designated keys on the keyboard with their left and right index fingers ("f" and "j" keys). After both questions were answered, a jittered silent inter-trial interval (ITI) ranging from 16 to 20 seconds followed, allowing the hemodynamic response to return to baseline.

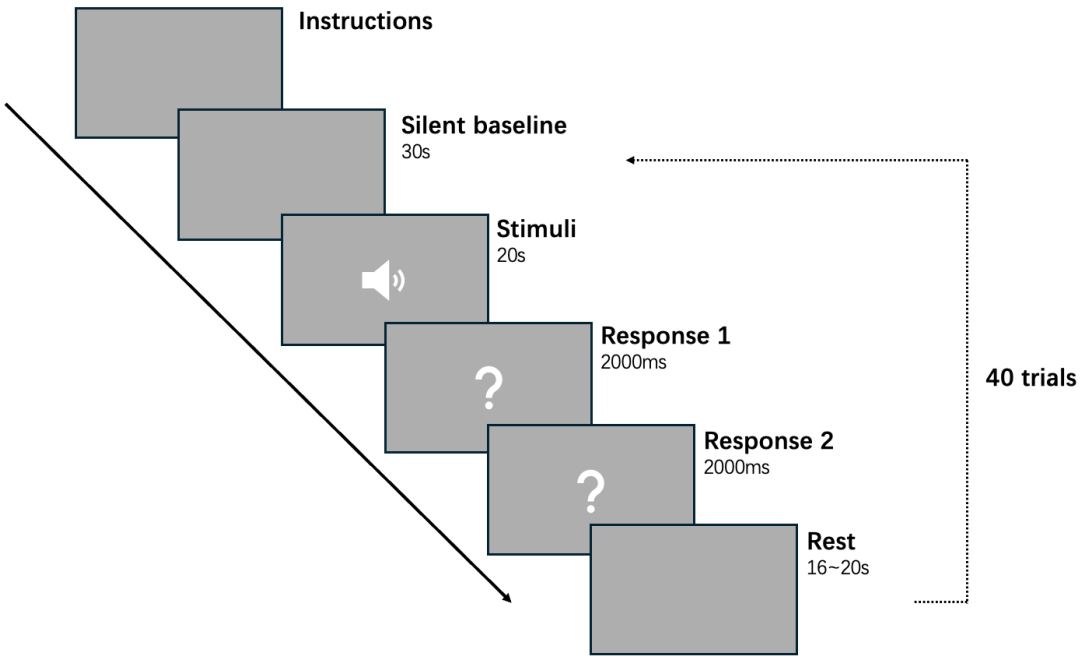


Figure 5.1 Procedure of fNIRS study

The task consisted of 40 unique trials in total (4 vocalization types: song, solfège, spoken lyrics, spoken story × 10 unique items). As hypothesized, repetition of stimuli was intentionally avoided to minimize confounding effects of in-experiment learning

on participants' familiarity and recollection judgments. The presentation order of the 40 trials was fully randomized for each participant. The entire fNIRS task lasted approximately 25 minutes.

Following the completion of the fNIRS task, the cap was removed. Participants then completed a battery of computerized questionnaires, which included the Profile of Music Perception Skills (PROMS), the Goldsmiths Musical Sophistication Index (Gold-MSI), and a demographic questionnaire that included the Edinburgh Handedness Inventory. The entire experimental session, including setup and post-task questionnaires, lasted approximately 75 minutes. Upon completion, participants were debriefed and received financial compensation for their time.

5.2.4 fNIRS setting

Hemodynamic changes in the cerebral cortex were recorded using a continuous-wave NIRx NIRSport 2 functional near-infrared spectroscopy (fNIRS) system. The system employed two wavelengths of near-infrared light, 760 nm and 850 nm, and data were acquired at a sampling rate of 10.2 Hz using NIRx Arora software (NIRx Medical Technologies, 2021).

The fNIRS probe set consisted of 16 sources and 15 detectors, which were fitted into a cap positioned according to international 10-20 system landmarks. The standard source-detector separation for long-distance channels was maintained at around 30 mm. The montage also included eight short-distance channels with an 8-mm separation, designed to measure and regress out superficial, extra-cerebral signals (e.g., scalp blood flow) during data preprocessing.

The optode montage was designed and optimized using the fOLD (fNIRS Optodes' Location Decider) toolbox in MATLAB (Zimeo et al., 2018). The design was targeted

to provide maximal coverage of a set of predefined regions-of-interest (ROIs) implicated in language, music, and auditory processing from previous literature. Based on the AAL2 brain atlas (Rolls et al., 2015) for segmentation, the selected ROIs for the present study included the bilateral superior frontal gyrus (SFG), middle frontal gyrus (MFG), middle temporal gyrus (MTG), and superior temporal gyrus (STG). The final montage resulted in the creation of 41 long-distance channels and 8 short-distance channels, with the full layout depicted in Figure 14.

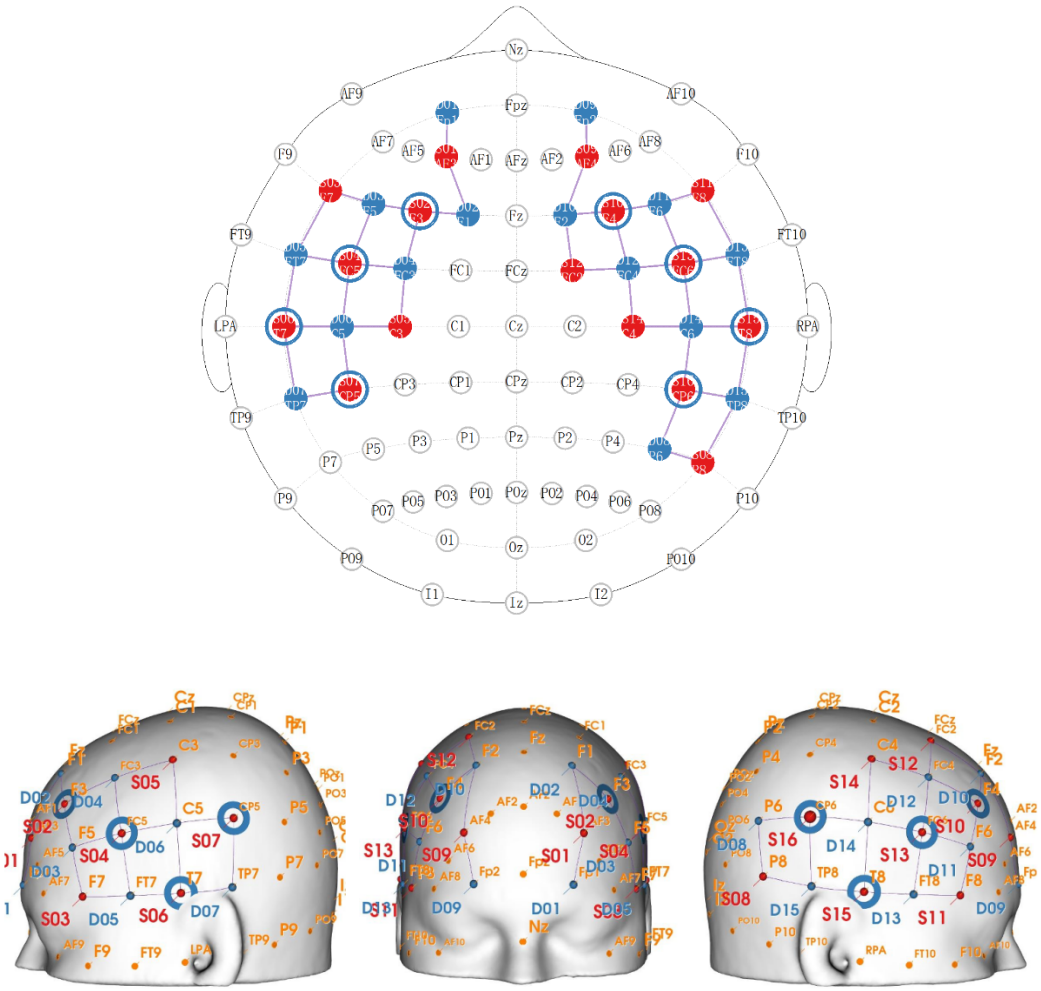


Figure 5.2 The design of fNIRS montage (upper: 2D version, bottom: 3D version)

To ensure that the hemodynamic signals analysed strictly corresponded to the targeted cortical structures, we performed a rigorous anatomical registration using the fOLD

(fNIRS Optodes' Location Decider) toolbox. Given the inherent anatomical variability in cortical folding (particularly in the temporal lobes), we utilized the LPBA40 (LONI Probabilistic Brain Atlas) for spatial mapping. Unlike deterministic atlases that define rigid boundaries, the LPBA40 atlas provides a probabilistic assessment of brain structures, making it highly suitable for fNIRS studies where surface-based measurements must be mapped onto volumetric cortical data.

Channel selection was governed by the Maximum Probability assignment strategy. Rather than applying a uniform, arbitrary percentage threshold which risks excluding boundary-spanning channels, each channel was assigned to the specific ROI with which it shared the highest cumulative probability of spatial overlap. This approach offers two distinct methodological advantages:

Optimization of Auditory Coverage: The boundary between the Superior Temporal Gyrus (STG) and the Middle Temporal Gyrus (MTG) is morphologically complex. The probabilistic assignment allowed us to effectively disentangle these regions, ensuring that channels capturing auditory processing (STG) were distinguished from those capturing semantic processing (MTG). As shown in Table 5.1, this method successfully identified bilateral coverage for both the STG and MTG, providing a robust basis for testing our dual-pathway hypotheses.

Validation of Montage Precision: The efficacy of this registration is evidenced by the specificity values reported in Table 5.1. For the prefrontal regions, which serve as the core of our executive function analysis, the spatial specificity was exceptionally high. The majority of channels in the Left and Right Middle Frontal Gyrus (MFG) and Inferior Frontal Gyrus (IFG) exhibited probability values exceeding 0.90 (e.g., Ch 1, 5, 26, 27, 29, 34). This confirms that the optode montage was accurately positioned over the targeted executive networks.

Consequently, even for channels located at cortical transition zones (e.g., Ch 10 in the Precentral Gyrus or Ch 32 in the STG), the Maximum Probability rule ensures that the data is attributed to its most anatomically plausible source, maximizing the retention of valid physiological signals while minimizing partial volume effects.

Table 5.1 Channels selected for the fNIRS analysis

Region of Interest (ROI)	**Hemisphere**	**Channel**	**Probability (Specificity)**
Precentral Gyrus	Left	14	0.843
		10	0.480
	Right	36	0.812
		40	0.675
Middle Frontal Gyrus (MFG)	Left	1	0.993
		2	0.979
		3	0.862
		4	0.899
		5	1.000
	Right	25	0.983
		26	0.996
		27	1.000
		29	1.000
		34	1.000
Inferior Frontal Gyrus (IFG)	Left	7	0.911
		9	1.000

	Right	31	0.990
		35	0.666
Superior Temporal Gyrus (STG)	Left	8	0.918
		17	0.705
	Right	43	0.930
		32	0.600
Middle Temporal Gyrus (MTG)	Left	16	0.934
		18	0.909
		21	0.633
	Right	44	0.896
		42	0.838
		48	0.796

Preprocessing was performed using the NIRS Toolbox in Matlab. To remove motion artifacts, we applied the Temporal Derivative Distribution Repair (TDDR) algorithm. The data were then band-pass filtered between 0.01 and 0.12 Hz to remove physiological noise and low-frequency drift.

For the statistical analysis, a subject-level General Linear Model (GLM) was applied. Short-separation channels were included as regressors in the GLM to account for systemic physiological interference. Beta values representing the hemodynamic response for each condition were extracted. Group-level analysis was subsequently performed to examine the effects of Stimulus Type (Song, Solfège, Lyrics, and Story) and Familiarity (Familiar vs. Unfamiliar).

5.3 Results

The analysis of fNIRS data examined the effects of Vocalization type (Song, Solfège, Spoken Lyrics, Spoken Story) and Familiarity (Familiar vs. Unfamiliar) on cortical hemodynamic responses. Linear mixed-effects models were fitted to oxygenated (HbO) and deoxygenated (HbR) hemoglobin data. All reported p values were corrected for multiple comparisons across channels ($p_{adj-global}$) or within specific regions of interest ($p_{adj-hier}$).

5.3.1 Main Effects of Vocalization and Familiarity

The analysis revealed a significant main effect of Vocalization on HbO concentration in the right prefrontal cortex, specifically in regions associated with cognitive control and attention. This was observed in both the right superior frontal gyrus (R-SFG), $F(3,522.94) = 4.97, p_{adj-global} = .025$, and the right middle frontal gyrus (R-MFG), $F(3,704.54) = 4.74, p_{adj-global} = .025$. Post-hoc contrasts confirmed that HbO

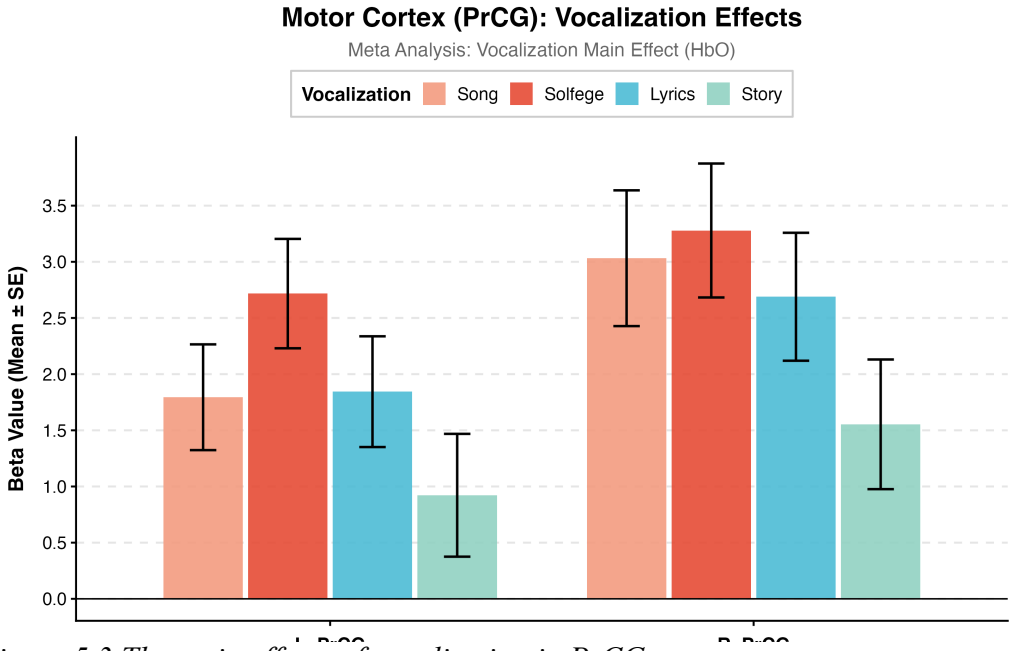


Figure 5.3 The main effects of vocalization in PrCG

activation was significantly greater for Song ($p_{adj-hier} = .012$), Solfège ($p_{adj-hier} = .012$), and Spoken Lyrics ($p_{adj-hier} = .041$) compared to the naturalistic Story condition. No significant main effect of Vocalization was found for HbR.

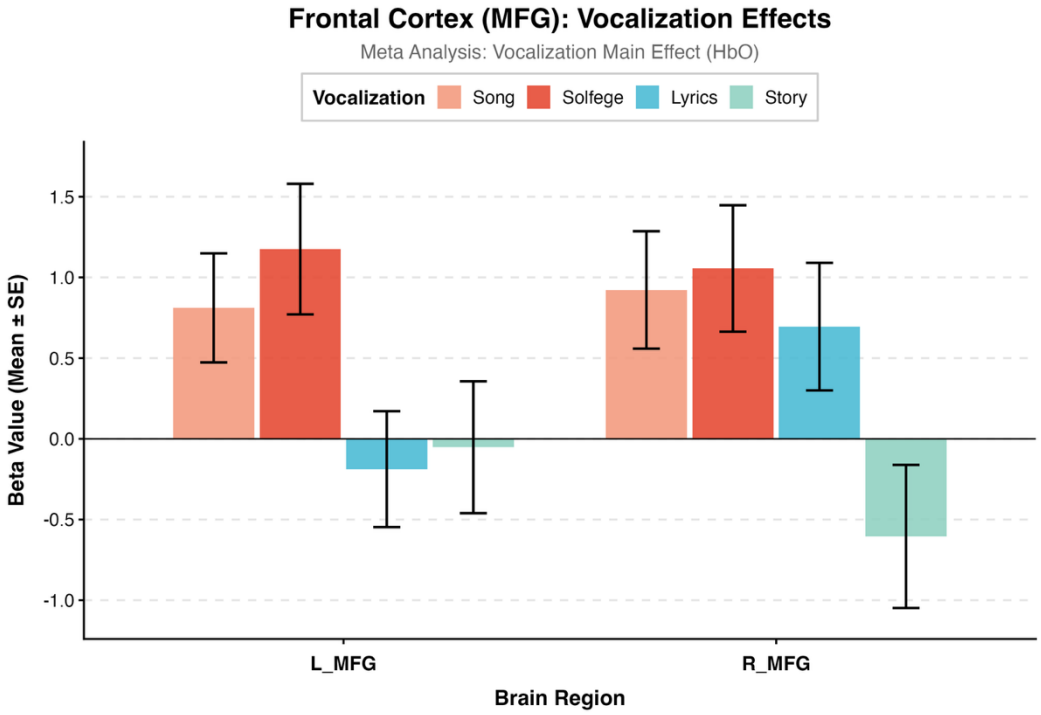


Figure 5.4 The main effect of vocalization in MFG

Regarding the main effect of Familiarity, significant modulation was found in hemodynamic responses, primarily in HbR concentration. In the R-SFG, familiar stimuli were associated with a significantly greater decrease in HbR than unfamiliar stimuli ($estimate = -0.62, F(1,527.50) = 10.25, p_{adj-global} = .026$). Conversely, in the left middle temporal gyrus (L-MTG), familiar stimuli were associated with a significant increase in HbR compared to unfamiliar stimuli ($estimate = 0.85, F(1,525.36) = 7.82, p_{adj-global} = .048$).

5.3.2 Interaction Effects between Vocalization and Familiarity

Crucially, a significant interaction between Vocalization and Familiarity was observed, revealing that the effect of familiarity was diametrically opposed for Song and Solfège.

Solfège (Familiarity decreases activation): For the solfège condition, familiarity was associated with a general decrease in activation, consistent with neural efficiency. HbO concentration in the right superior temporal gyrus (R-STG; $estimate = -2.08, t(164.55) = -2.28, p_{adj-hier} = .024$) and the right middle temporal gyrus (R-MTG; $estimate = -1.54, t(528.60) = -2.06, p_{adj-hier} = .040$) was significantly lower for familiar solfège compared to unfamiliar solfège. Similarly, the decrease in HbR in the R-MFG was significantly smaller for familiar solfège than for unfamiliar solfège ($estimate = -0.79, t(710.85) = -2.66, p_{adj-hier} = .008$).

Novelty Effect: Additionally, when solfège was unfamiliar, it elicited significantly greater HbO activation in the left precentral gyrus (L-PrCG) compared to the Story baseline ($t(346.11) = 3.22, p_{adj-hier} = .046$).

Song (Familiarity increases activation): For the song condition, the effect of familiarity was the opposite. HbO concentration in the Left Middle Frontal Gyrus (L-MFG) was significantly greater for familiar songs compared to unfamiliar songs ($estimate = 1.70, t(528.47) = 1.96, p_{adj-hier} = .050$).

Shared Effects (Song and Lyrics): Further analysis of the interaction showed that only Familiar Songs and Familiar Lyrics elicited significantly heightened HbO activation in the R-MFG relative to the spoken story baseline ($p_{adj-hier} = .027$ for both).

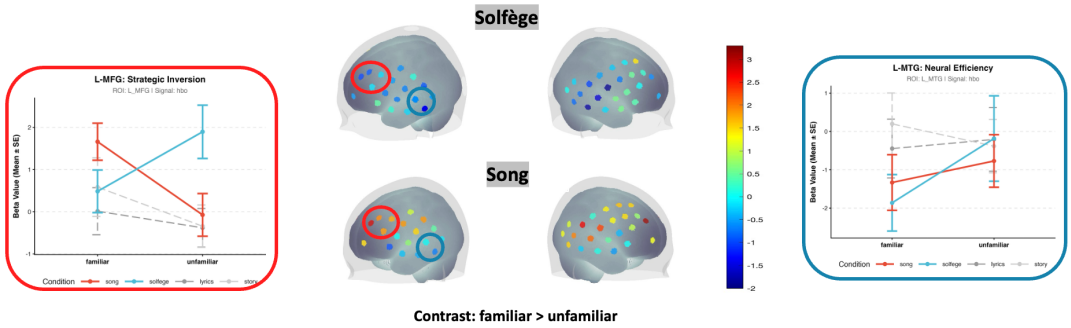


Figure 5.5 Opposing activation patterns for Song versus Solfège in left fronto-temporal regions

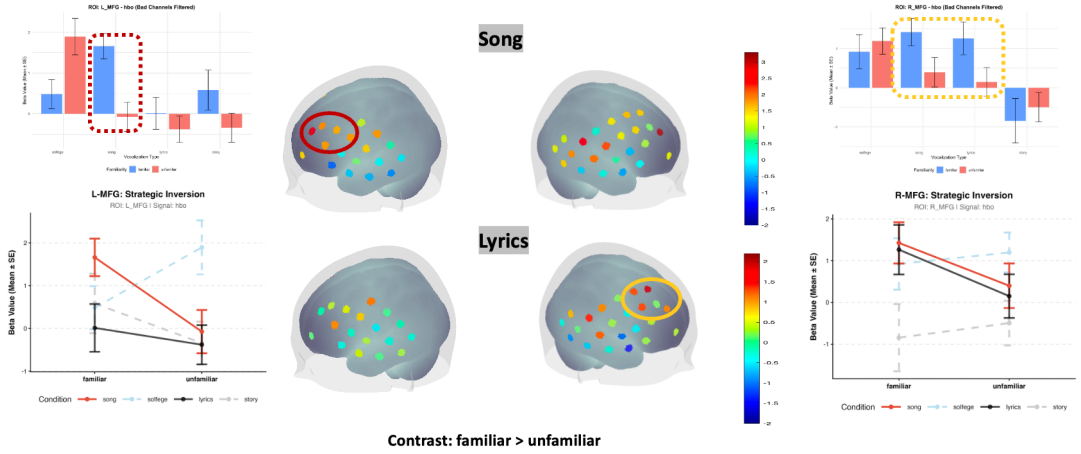


Figure 5.6 Song and Lyrics share the same directional trend in the bilateral MFG

5.4 Discussions

The present fNIRS study investigated the neural correlates of song perception, focusing on how the brain integrates lyrical and melodic information and how this process is modulated by familiarity. The results revealed a critical dissociation in the processing of linguistic-meaningful song versus linguistic-meaningless melody (solfège). These findings challenge models of simple additivity and instead support a Dual-Pathway Model, where familiarity acts as a switch transitioning the brain from auditory-motor simulation (for solfège) to high-order semantic integration (for song).

5.4.1 Structural Complexity and General Sensitivity

A key baseline finding was the main effect of vocalization in the Right Prefrontal Cortex (R-PFC). Compared to the naturalistic Story baseline, all three structured vocalizations (Song, Solfège, and Spoken Lyrics) elicited greater neural activation in the R-SFG and R-MFG. This suggests that vocalizations with a predictable, patterned structure — whether musical or metrical — demand greater top-down monitoring and working memory resources than the more automatic flow of narrative speech. This establishes that the brain is generally sensitive to the structural complexity of a vocal signal, independent of its specific content.

5.4.2 Solfège: Automation via Skill Acquisition

The central finding of this experiment lies in the interaction effects. For solfège, the results confirm that processing is essentially a skill acquisition process. Unfamiliar solfège recruited the Left Precentral Gyrus (L-PrCG), a region central to the dorsal auditory-motor stream. This suggests that when decoding a novel melodic sequence with arbitrary syllables, the brain engages in predictive simulation or subvocal rehearsal to track the pitch-syllable mapping.

However, once the stimulus becomes familiar, this processing becomes automated. We observed a significant drop in activation in right temporal and prefrontal regions for familiar solfège. This pattern is consistent with the Neural Efficiency Hypothesis, where practice leads to a reduction in the neural resources required to execute a task. The brain shifts from an effortful, rule-based decoding strategy to a more efficient, automatized response.

5.4.3 Song: Active Retrieval and Integration

In striking contrast, familiarity with a meaningful song did not lead to neural efficiency but to enhanced neural engagement. This was most evident in the increased activation in the Left Middle Frontal Gyrus (L-MFG) for familiar songs. This suggests that for familiar songs, the brain is not saving effort but is engaging in a deeper cognitive operation.

We propose that in the context of song, the L-MFG acts as a conductor or central executive. Rather than merely storing musical traces, the L-MFG actively searches for, selects, and maintains semantic (lyrics) and melodic information from long-term memory. This top-down retrieval effort is absent in unfamiliar songs, where no prior memory traces exist to be retrieved. Therefore, the heightened activation reflects the cognitive load of binding: integrating the incoming auditory input with a rich, pre-compiled multi-modal schema (lyrics, melody, and emotional context).

5.4.4 The Role of Linguistic Meaning

The specificity of this retrieval mechanism is further supported by the shared pattern between Song and Spoken Lyrics. Only familiar songs and familiar lyrics elicited heightened activation in the R-MFG relative to the Story baseline. This convergence strongly suggests that the neural enhancement driven by familiarity is primarily linked to the processing of meaningful linguistic content. The presence of semantics fundamentally alters how the melody is processed; it prevents the brain from treating the signal as a simple acoustic pattern (like solfège) and instead mandates a resource-intensive process of semantic access and integration.

In conclusion, this study provides neuroimaging evidence for a hierarchical mechanism in Cantonese song perception. We demonstrate that familiarity is not a monolithic

construct; it exerts opposing effects on the prefrontal cortex depending on the nature of the stimulus. It facilitates automatization and neural efficiency in rule-based tasks (solfège) but triggers active memory retrieval and increased neural load in naturalistic songs. This confirms that song perception is a dynamic process where the presence of meaningful linguistic content shifts the brain from a mode of auditory-motor simulation to one of high-order cognitive integration.

6 General Discussion

The current project was designed to investigate the complex relationship between speech and music perception, using song as a critical, ecologically valid stimulus to bridge the two domains. Across three experiments, we examined this relationship at the perceptual, cognitive, and neural levels.

In Experiment 1, we examined the perceptual nature of the speech-song boundary using a carefully synthesized acoustic continuum that preserved linguistic content and rhythmic patterns while varying only the acoustic features distinguishing speaking from singing. Results revealed a non-CP pattern, with gradually increasing discrimination sensitivity along the speech-song continuum, and this pattern was influenced by participants' musical competence indicated by PROMS scores.

Experiment 2 investigated how different auditory contexts affect lexical tone normalization, demonstrating a hierarchical pattern of context effects (speech > song > solfège/instrumental music) and showing that musical training modulates these effects by enhancing the musical interpretation of song contexts. Notably, song and solfège contexts differed not in acoustic properties (both involved singing rather than speaking) but in content (lyrics versus solmization syllables).

Finally, Experiment 3 employed fNIRS to explore the neural processing of different vocalizations with a specific attention on song, finding distinct neural strategies for four vocalization types with two levels of familiarity, with songs of higher familiarity eliciting deeper prefrontal engagement compared to the neural efficiency observed for familiar melodies (solfège).

Collectively, these findings suggest that speech and music processing exist along a functional continuum rather than as entirely separate systems, with song perception serving as a critical window into their interaction. Below, we discuss these findings in

relation to existing theoretical frameworks and propose an integrated model that mainly accounts for the continuity between speech and music processing.

6.1 The Speech-Song Continuum: From a Continuous Perceptual Space to a Distinction Between Cognitive Domains

The first experiment addressed the fundamental nature of the acoustic perceptual boundary between speech and song. Our key finding was that the perception of a continuous acoustic transition from speech to song is non-categorical. Listeners did not exhibit the discrimination accuracy peak that corresponds with the sharp identification boundary which would signify the presence of two discrete perceptual categories. Instead, the results showed a pattern of continuously increasing discrimination accuracy as stimuli became more song-like.

This finding of a continuous perceptual space seems to contrast with recent neuroimaging research providing strong evidence for a neural separation of speech and music (e.g., Albouy et al., 2024). However, we argue that these findings are not contradictory but are, in fact, complementary. Albouy et al. (2024) cleverly separated a song into its constituent temporal and spectral dimensions, demonstrating that the brain possesses functionally lateralized systems for processing the core acoustic features typically associated with lyrics and melody, respectively. Our study, in contrast, did not separate these dimensions but created a true acoustic continuum where speech and song features were gradually blended. Our results therefore suggest that while the brain has specialized systems for analysing different acoustic components, when those components are integrated into a single, ambiguous signal, a rigid categorical boundary does not emerge at the perceptual level.

The interpretation of this continuous perceptual pattern is supported by phenomenological evidence like the speech-song illusion (Deutsch et al., 2011), which demonstrates that the boundary is fluid. We propose that the continuously increasing discrimination accuracy observed in our data reflects a gradual shift in cognitive listening strategy. The primary goal of speech perception is to achieve perceptual constancy, a process that requires the listener to ignore irrelevant acoustic variations in order to extract a stable, invariant linguistic label (Kuhl, 2004). In contrast, the aesthetic appreciation of music often requires listeners to attend to these subtle variations in performance, such as timbre and expressive nuance, as a source of meaning and emotion (Gabrielsson, 2003). The pattern in our data is therefore consistent with a gradual transition from a speech-oriented mode that normalizes variance to a music-oriented mode that values it.

This functional divergence leads to a deeper theoretical proposal regarding the relationship between speech and song. We argue that the absence of a simple categorical boundary is because the distinction is not between two adjacent perceptual categories, but between two distinct, albeit highly interactive, cognitive domains. Perceptual categories are often defined by a boundary along a limited set of acoustic dimensions. Cognitive domains, such as speech and music, are vast computational systems, each with unique rule structures, goals, and neural substrates. Therefore, the gradual transition our listeners perceived may not reflect a simple switch in a feature detector, but a more complex, graded shift in the dominant processing mode from the linguistic domain to the musical domain. This reframes the speech-song continuum not merely as an acoustic space, but as a landscape of shifting cognitive engagement between two of the brain's most sophisticated expert systems.

6.2 Functional Divergence and the Cumulative Cue Framework

While the first experiment established a continuous perceptual space at the acoustic level, the second experiment demonstrated that a clear functional hierarchy emerges when a specific linguistic mechanism, namely, contextual normalization, is engaged. The results confirmed our central hypothesis: the normalization effect followed a distinct, graded pattern (Speech > Song > Solfège ≈ Instrumental Music).

This result is best understood within the cumulative cue framework, which posits that speech normalization is not an all-or-nothing process but a graded one, where the strength of the effect increases as more linguistically relevant cues are present in the signal. For instance, C. Zhang & Chen (2016) systematically demonstrated this additive principle by showing that the normalization effect grew stronger as the context evolved from non-speech sounds to include phonetic, then phonological, and finally semantic information. Our study does not refute this model but rather extends it in two critical ways by using vocal music to disentangle vocal, musical, and linguistic cues.

First, our findings support the additive nature of linguistically functional cues in speech normalization. The critical dissociation was between the song and solfège contexts. Both stimuli are vocal and share a musical structure, yet only the song context induced a robust normalization effect. This cannot be attributed to a simple lack of semantic content, as both our own results and previous studies (e.g., C. Zhang & Chen, 2016) have shown that meaningless speech (containing phonologically legal but no semantic meanings) can elicit a strong, speech-like normalization effect. Instead, the crucial difference lies in the functional interpretation of the syllables and the structure in which they are embedded. The syllables in a song, while set to music, are processed as components of the linguistic system intended to convey a message. In contrast, the syllables in the solfège context (e.g., *do, re, mi*), while phonetically complex, are

primarily interpreted within a musical system to denote scale degrees and harmonic function. This musical function, combined with the rigid musical structure that lacks naturalistic speech prosody, appears to prevent the signal from fully engaging the specialized speech normalization mechanism. This demonstrates the primacy of linguistic function: for a vocal cue to be treated as "speech-like" by this system, it must be perceived as operating within a linguistic framework, not just be phonetically complex.

Second, and most critically, our study demonstrates that the cumulative cue framework must account not only for facilitatory linguistic cues but also for competing or inhibitory non-linguistic cues. This is evidenced by the finding that the normalization effect for the song context was significantly weaker than for natural speech contexts. While the song contains the necessary vocal and linguistic cues to trigger normalization, its musical structure (e.g., stable pitch, metrical rhythm) is a feature of nonspeech. The intermediate effect elicited by song suggests a perceptual compromise, where the presence of this musical structure actively modulates or reduces the overall strength of the effect. Therefore, the strength of the normalization effect appears to be the outcome of a dynamic weighting of both facilitatory linguistic cues and competing musical-structural cues present in a complex vocal signal. This provides a more comprehensive understanding of the cumulative cue model.

6.3 Top-Down Modulation and the Transitional Nature of Song

Experiment 3 extended our investigation from perceptual and functional categorization to the realm of higher-order cognitive processing. By introducing familiarity as a variable, we examined how long-term memory and semantic access modulate the neural processing of Song, Solfège, and Lyrics. The fNIRS results revealed that the brain's

categorization of these stimuli is not rigidly determined by their acoustic properties but is dynamically reconfigured by top-down cognitive demands.

The most compelling evidence for the transitional nature of song lies in the dissociation between acoustic category and functional activation patterns under the modulation of familiarity.

First, we observed a divergent pattern within the vocal music category**.** Although Song and Solfège share identical melodic and rhythmic structures (both being sung), their neural responses to familiarity were diametrically opposed. High familiarity with Solfège elicited a pattern of neural efficiency, suggesting that when the brain recognizes a purely structural musical signal, it streamlines processing resources. In contrast, high familiarity with Song elicited neural enhancement (increased activation). This indicates that vocal music is not a monolithic processing entity; the presence of semantic content in Song fundamentally alters the neural pathway, overriding the efficiency mechanism typically seen in melodic processing.

Second, and critically, we observed a functional convergence between Song and Lyrics. Despite their distinct acoustic classifications (music vs. speech), these two conditions exhibited a parallel direction of regulation under the influence of familiarity. Both familiar Song and familiar Lyrics triggered enhanced prefrontal engagement. This finding implies that when a stimulus is familiar and meaningful, the brain prioritizes its semantic function over its acoustic form. The Song condition, therefore, behaves like Music when unfamiliar (relying on structural analysis) but shifts toward Language when familiar (recruiting semantic retrieval networks).

This pattern of results supports the hypothesis that Song occupies a transitional zone between music and speech. It is not merely a static superposition of melody and words. Instead, it functions as a flexible medium where the balance between musical and

linguistic processing is modulated by top-down factors. The identity of a song is fluid; it shifts toward the linguistic end of the continuum when memory and meaning are actively engaged, thereby utilizing resources shared with spoken language.

6.4 Toward a Multi-Stage Interactive Model of Vocal Perception

Collectively, the three experiments in this thesis paint a cohesive, multi-stage picture of how the human brain navigates the continuum between speech and music. By synthesizing the findings from fine-grained psychophysics (Experiment 1), contextual gating (Experiment 2), and hemodynamic modulation (Experiment 3), we propose a Multi-Stage Interactive Model of vocal perception.

Stage 1: Acoustic-Perceptual Continuity (Bottom-Up)

At the initial stage, as demonstrated in Experiment 1, the processing of vocal signals is non-categorical. The brain analyzes spectro-temporal features without imposing rigid categorical boundaries. At this level, music and speech are not distinct modules but rather poles on a sensory continuum. This explains why objective perceptual skills (PROMS) were the strongest predictor of performance in our initial discrimination tasks: the system is primarily engaged in fine-grained acoustic analysis.

Stage 2: Functional Gating and Divergence

At a subsequent stage, a functional divergence occurs. As shown in Experiment 2, once the brain detects specific acoustic cues (or contextual frames) that signal a linguistic or musical intent, distinct processing streams are engaged. This stage acts as a gatekeeper, directing resources based on the perceived function of the signal. Here, the binary status of professional musical training becomes relevant, likely by providing more refined templates for this categorization process, thus influencing how the signal is normalized.

Stage 3: Higher-Order Integration and Modulation (Top-Down)

Finally, Experiment 3 reveals that these streams are not encapsulated. At the highest level, processing is non-additive and profoundly modulated by top-down factors such as memory and familiarity. This stage is where the system exhibits maximum flexibility. As evidenced by the fNIRS data, familiarity can override low-level acoustic categories. The fact that Song aligned with Lyrics (Speech) rather than Solfège (Vocal Music) under high-familiarity conditions indicates that higher-order cognition prioritizes semantic access. When a song is familiar, the music pathway and language pathway do not merely run in parallel; they interact synergistically, with semantic memory pulling the processing of Song toward the linguistic domain.

To conclude, this model reconciles previous theoretical conflicts by positing that modularity and shared resources are not mutually exclusive but are operational at different processing stages. The brain utilizes specialized resources for structural analysis (supporting the modular view of Peretz & Coltheart, 2003) but integrates them into a shared semantic and executive network when the stimulus is meaningful and familiar (supporting the shared resource view of Patel, 2003). Ultimately, this research uses Song as a natural probe to unify our understanding of auditory cognition. The findings demonstrate that the relationship between music and language is best described not as a dichotomy, but as a dynamic continuity. This continuity is anchored by acoustic features at the bottom but is fluidly navigated at the top, driven by the listener's experience, memory, and the search for meaning.

7 Appendices

7.1 Supplementary Tables for Statistical Analyses

Table 7.1 Estimated Sample Size Requirements for Different Analysis Types and Effect Sizes in Experiment 1

Analysis Type	Effect Size	Required Sample Size
Identification task - boundary position (medium effect)	$f^2 = 0.15$	55
Identification task - boundary position (large effect)	$f^2 = 0.35$	25
Identification task - boundary slope (medium effect)	$f^2 = 0.15$	55
Identification task - boundary slope (large effect)	$f^2 = 0.35$	25
Discrimination task - accuracy (medium effect)	$d = 0.51$	28
Discrimination task - accuracy (large effect)	$d = 0.8$	52
Pair level effect (medium effect)	$d = 0.5$	34
Pair level effect (large effect)	$d = 0.8$	15
Interaction effect (small effect)	$f^2 = 0.02$	550
Interaction effect (medium effect)	$f^2 = 0.15$	77

Table 7.2 Model Selection Results for Linear Mixed-Effects Models Predicting the Identification Boundary

Model	Predictors	*df*	logLik	AICc	ΔAICc	*w*

model_b2	PROMS_z	4	-1515.83	3039.7	0	0.578
model_b6	trn + PROMS_z	5	-1515.85	3041.8	2.08	0.205
model_b8	PROMS_z + Gold_MSI_z	5	-1516.8	3043.7	3.97	0.079
model_b13	PROMS_z + each_f_z	5	-1517.35	3044.8	5.08	0.046
model_b14	PROMS_z + meanf_z	5	-1517.49	3045.1	5.36	0.04
model_b9	trn + PROMS_z + Gold_MSI_z	6	-1516.63	3045.4	5.68	0.034
model_b1	trn	4	-1519.75	3047.5	7.84	0.011
model_b7	trn + Gold_MSI_z	5	-1520.32	3050.7	11.01	0.002
model_b3	Gold_MSI_z	4	-1521.68	3051.4	11.71	0.002
model_b11	trn + each_f_z	5	-1521.04	3052.2	12.45	0.001
model_b12	trn + meanf_z	5	-1521.41	3052.9	13.19	0.001
model_b17	All predictors	8	-1518.95	3054.1	14.4	< .001
model_b15	Gold_MSI_z + each_f_z	5	-1523.04	3056.2	16.45	< .001
model_b16	Gold_MSI_z + meanf_z	5	-1523.35	3056.8	17.07	< .001
model_b4	each_f_z	4	-1524.84	3057.7	18.04	< .001
model_b5	meanf_z	4	-1525.2	3058.5	18.75	< .001
model_b10	each_f_z + meanf_z	5	-1525.41	3060.9	21.19	< .001

Note. Models are ranked by AICc. *df* = number of parameters, logLik = log-Likelihood, AICc = Akaike Information Criterion (corrected), ΔAICc = difference in AICc from the best model, *w* = Akaike weight. Predictors: trn = musical training, PROMS_z =

PROMS score, Gold_MSI_z = Goldsmiths Musical Sophistication Index score, each_f_z & meanf_z = measures of song familiarity.

Table 7.3 Model Selection Results for Linear Mixed-Effects Models Predicting the Identification Slope

Model	Predictors	*df*	logLik	AICc	ΔAICc	*w*
model_s2	PROMS_z	4	-345.07	698.2	0	0.559
model_s14	PROMS_z + meanf_z	5	-345.16	700.4	2.2	0.186
model_s3	Gold_MSI_z	4	-346.88	701.8	3.62	0.092
model_s8	PROMS_z + Gold_MSI_z	5	-346.84	703.8	5.56	0.035
model_s6	trn + PROMS_z	5	-346.93	703.9	5.75	0.032
model_s16	Gold_MSI_z + meanf_z	5	-346.97	704	5.81	0.031
model_s13	PROMS_z + each_f_z	5	-347.17	704.4	6.23	0.025
model_s5	meanf_z	4	-348.9	705.9	7.66	0.012
model_s1	trn	4	-349.23	706.5	8.32	0.009
model_s7	trn + Gold_MSI_z	5	-348.62	707.3	9.11	0.006
model_s15	Gold_MSI_z + each_f_z	5	-348.72	707.5	9.32	0.005
model_s9	trn + PROMS_z + Gold_MSI_z	6	-347.99	708.1	9.89	0.004
model_s12	trn + meanf_z	5	-349.32	708.7	10.53	0.003
model_s4	each_f_z	4	-350.52	709.1	10.9	0.002
model_s11	trn + each_f_z	5	-350.91	711.9	13.71	0.001

model_s10	each_f_z + meanf_z	5	-351.93	713.9	15.74	< .001
model_s17	All predictors	8	-351.08	718.4	20.16	< .001

Note. Models are ranked by AICc. See notes for Table 2 for abbreviation definitions.

Table 7.4 Hierarchical Model Comparison Results for Predicting Discrimination Accuracy (ACC)

Model Step	Predictors Added	npar	AIC	BIC	logLik	Deviance	χ^2	df	p
1	Base Model (Main Effects)	9	-3228.5	-3178.1	1623.2	-3246.5			
2	+ Pair_n × Gold-MSI	10	-3226.7	-3170.8	1623.4	-3246.7	0.25	1	0.614
3	+ Pair_n × PROMS	10	-3230.2	-3174.3	1625.1	-3250.2	3.53	*	*
4	+ Pair_n × Training	10	-3240.6	-3184.7	1630.3	-3260.6	10.38	*	*
5	… (other single interactions)								
6	All Interactions	13	-3251	-3178.4	1638.5	-3277	29.06	3	< .001
7	+ Additional Term 1	14	-3249.1	-3170.8	1638.6	-3277.1	0.06	1	0.801

8	+ Additional Term	2	16	-3246.5	-3157.1	1639.3	-3278.5	1.4	2	0.497

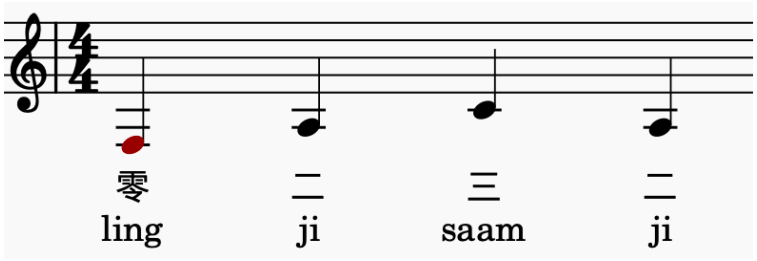
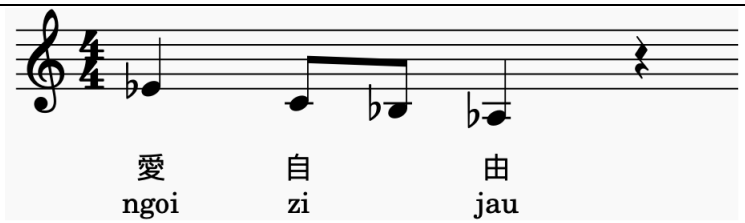
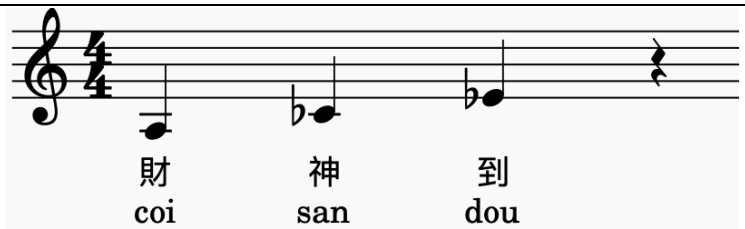
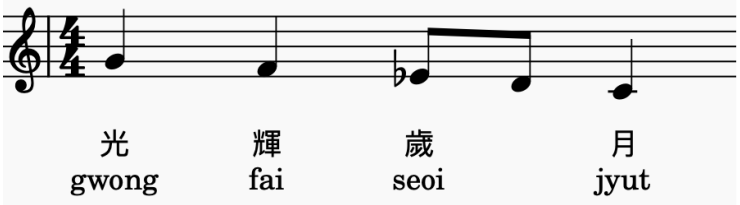
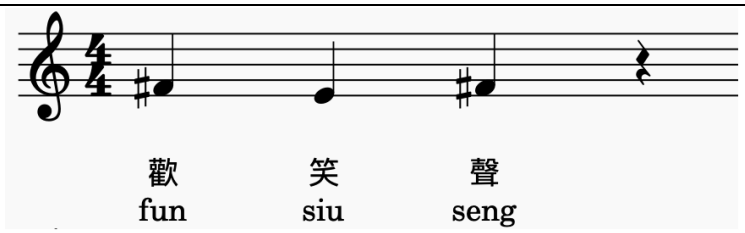
Note. Abridged table showing key steps of the hierarchical model comparison. Step 1 includes main effects. Steps 2-5 tested single interaction additions. Step 6 represents the selected final model, which included all theoretically motivated two-way interactions and showed a significant improvement over simpler models. Subsequent steps (7-8) showed that adding further terms did not improve the model. The model from Step 6 was analysed in detail.

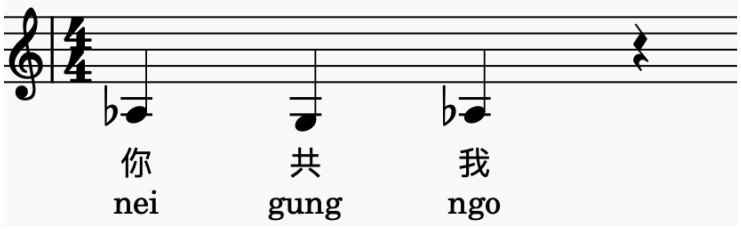
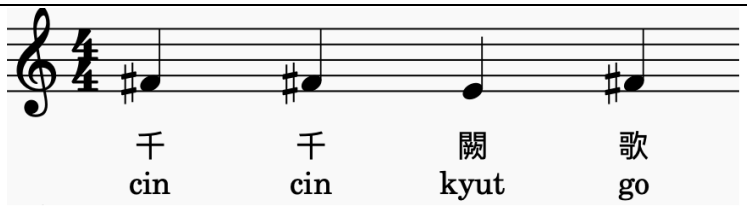
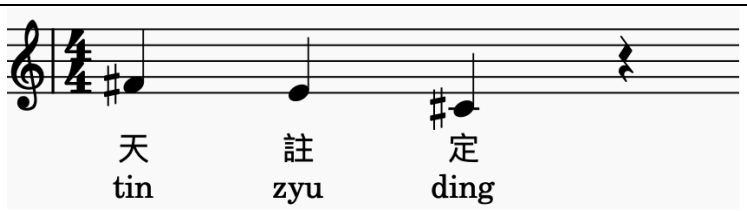
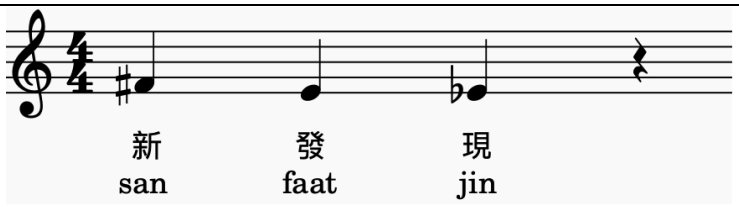
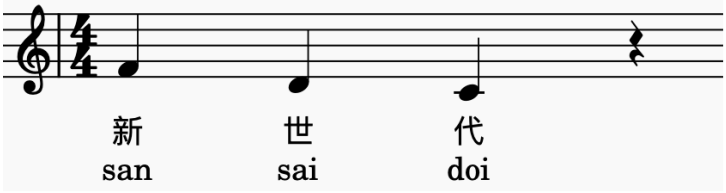
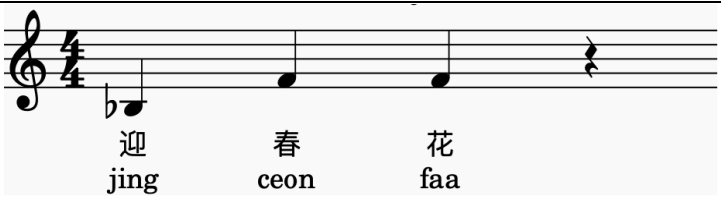
The provided table had missing Chi-squared test results for some steps, indicated by *.

7.2 Scores of musical materials

7.2.1 Materials for Experiment 1

Table 7.5 Eleven song excerpts for continuum synthesis (Experiment 1)

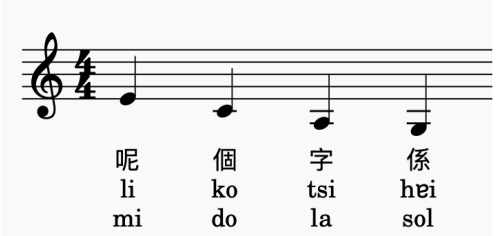
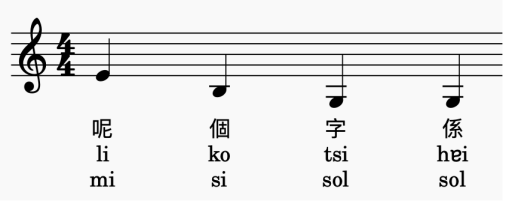
Song ID	Musical scores
0232	
AZY	
CSD	
GHSY	
HXS	

NGW	
QQQG	
TZD	
XFX	
XSD	
YCH	

7.2.2 Materials for Experiment 2

Table 7.6 Melodies of music contexts in Experiment 2

Singer	Melodies

F1 (soprano)	
F2 (alto)	

Note. The melody is presented in three contexts: instrumental music, song, and solfege. The lyrics are displayed in three lines: the first line shows the song lyrics, the second line provides the corresponding *Jyutping* pronunciation, and the third line presents the solfege syllables for the solfege context.

7.2.3 Sheet Music Referenced in Experiment 3

The following sheet music was referenced during the recording process of Experiment 3. Some of the sheet music is in simplified notation, and the key was adjusted to match the vocal range of the participants, meaning the performance was not in the original key of the song.

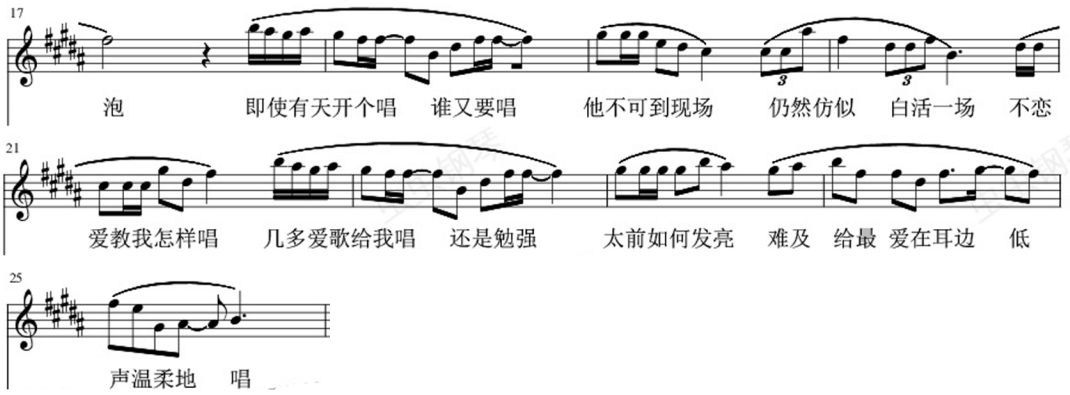


Figure 7.1 "Diva... Ah Hey" (下一站天后). Source: https://www.gangqinpu.com/

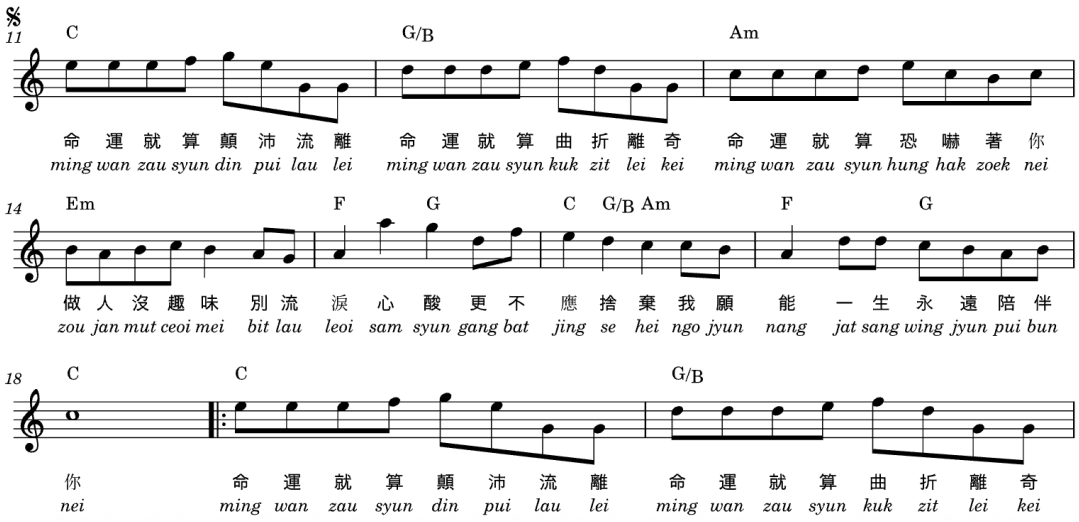
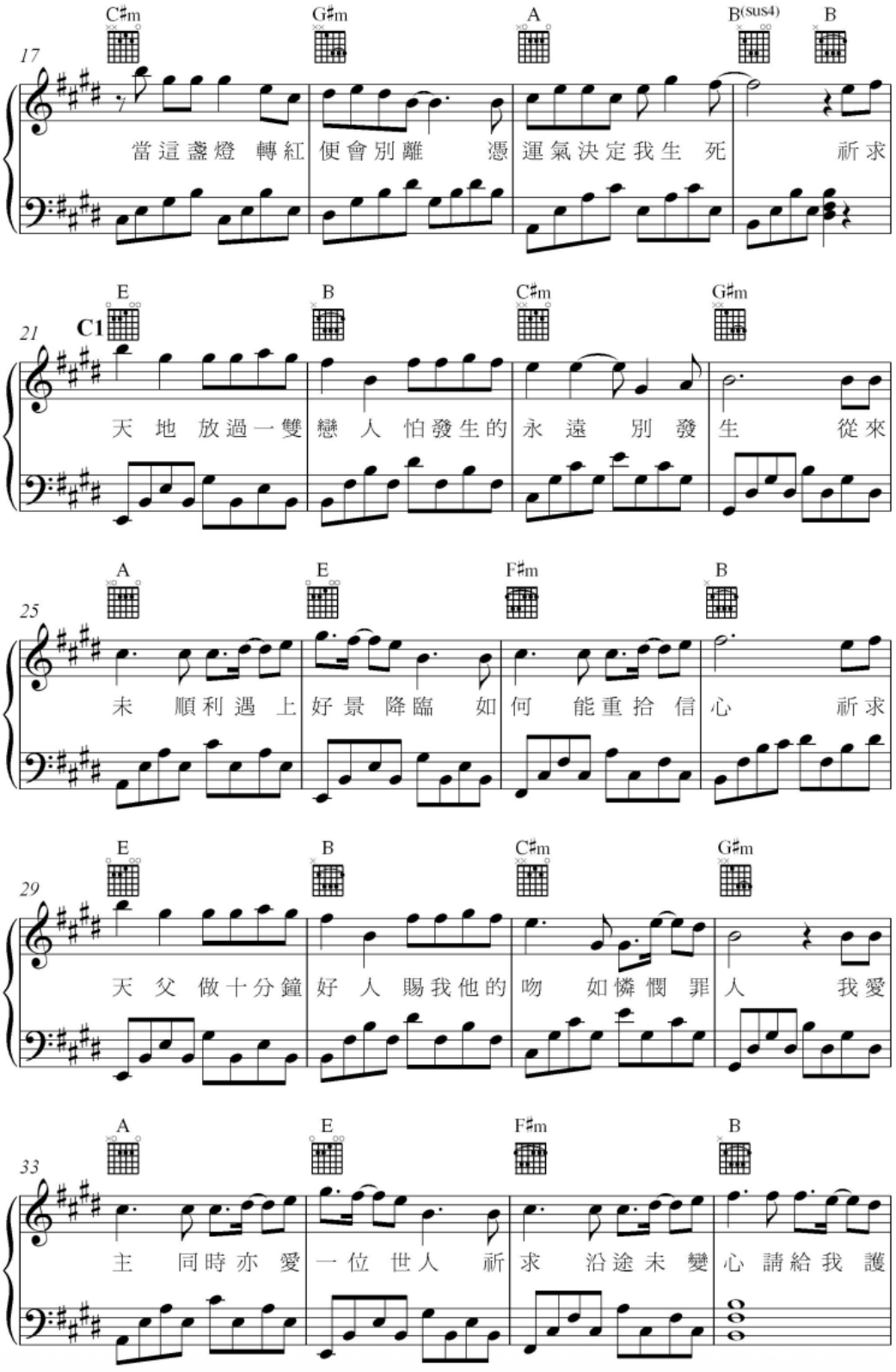


Figure 7.2 "Red Sun" (红日). Source: musescore.com


C♯m G♯m A B(sus4) B
17
當這盞燈 轉紅 便會別離 憑運氣決定我生 死 祈求
E B C♯m G♯m
21
C1
天 地 放過一雙 戀 人 怕發生的 永 遠 別 發 生 從來
A E F♯m B
25
未 順利遇 上 好景降臨 如何 能重拾 信 心 祈求
E B C♯m G♯m
29
天 父 做十分鐘 好 人 賜我他的 吻 如憐憫 罪 人 我愛
A E F♯m B
33
主 同時亦 愛 一位 世人 祈求 沿途未 變 心 請給我 護

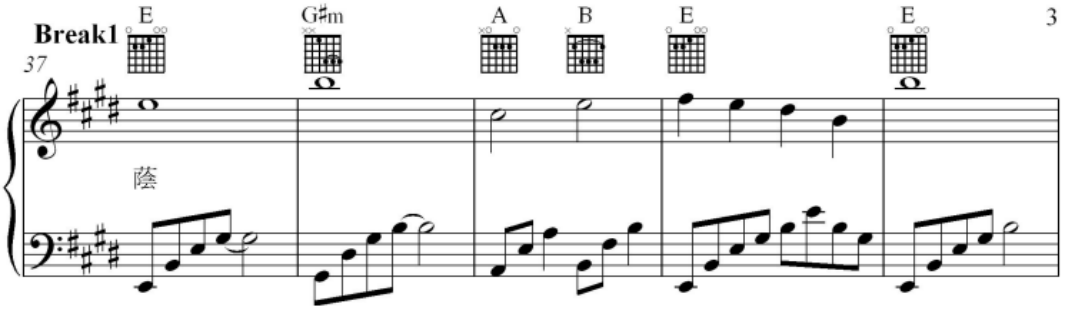


Figure 7.3 "A Maiden's Prayer" (少女的祈祷). Source: https://www.poppiano.org/

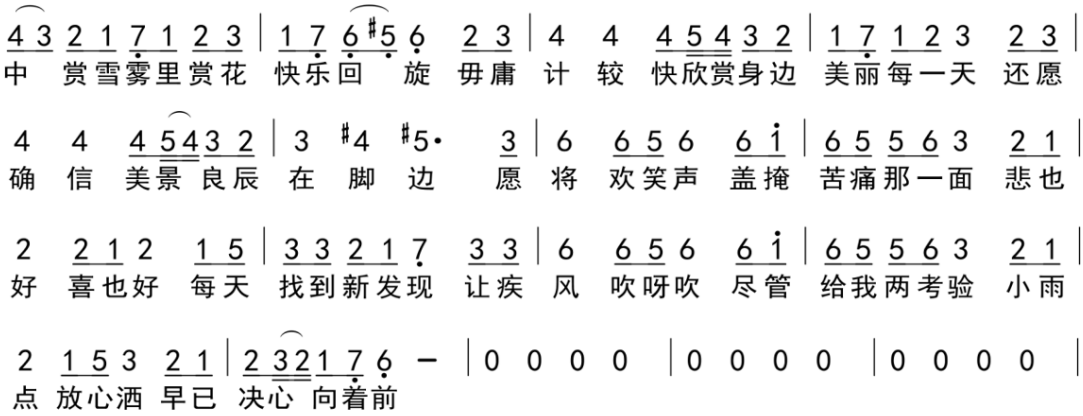


Figure 7.4 "Strolling Through Life's Path" (漫步人生路). Source: https://www.poppiano.org/

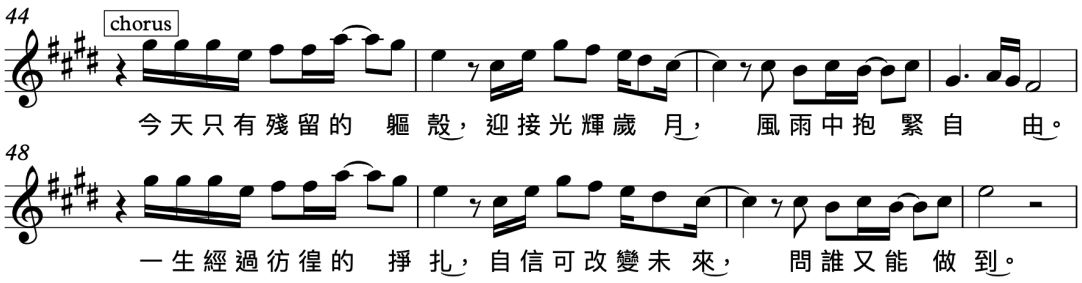


Figure 7.5 "Glorious Days" (光辉岁月). Source: https://www.poppiano.org/

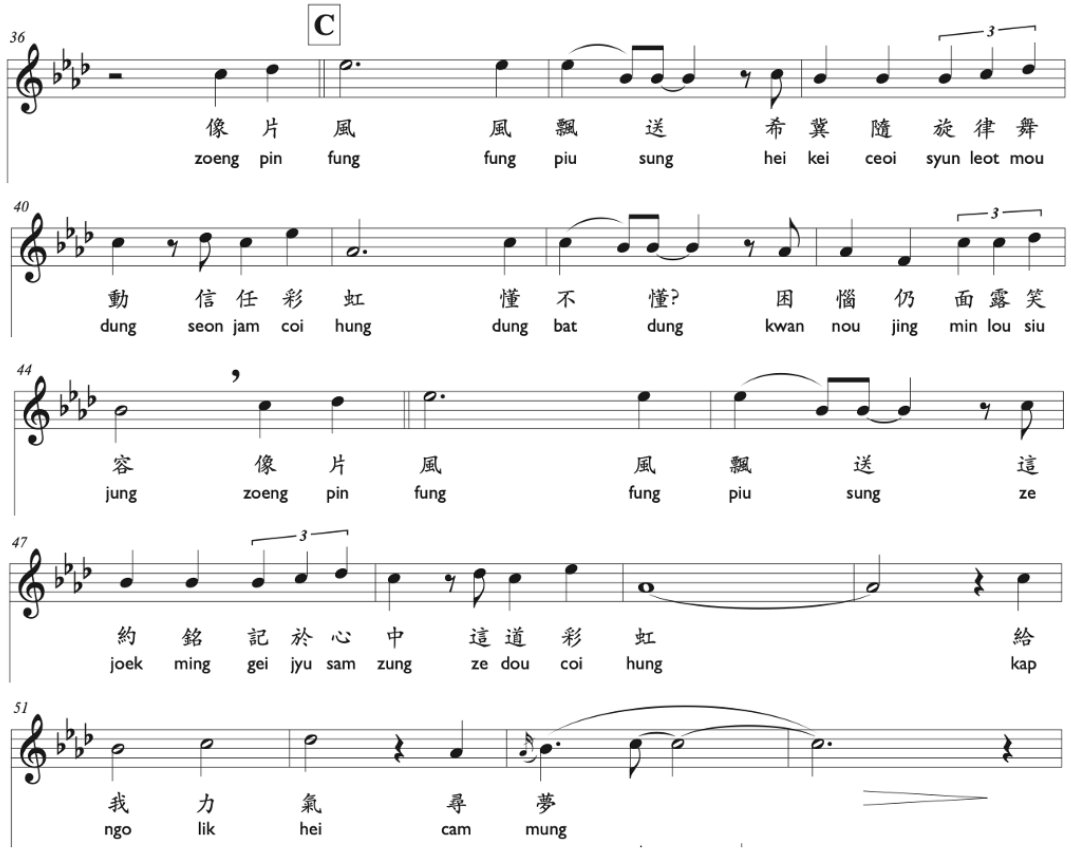


Figure 7.6 "Trusting the Rainbow" (信任彩虹). Authorized by the composer.

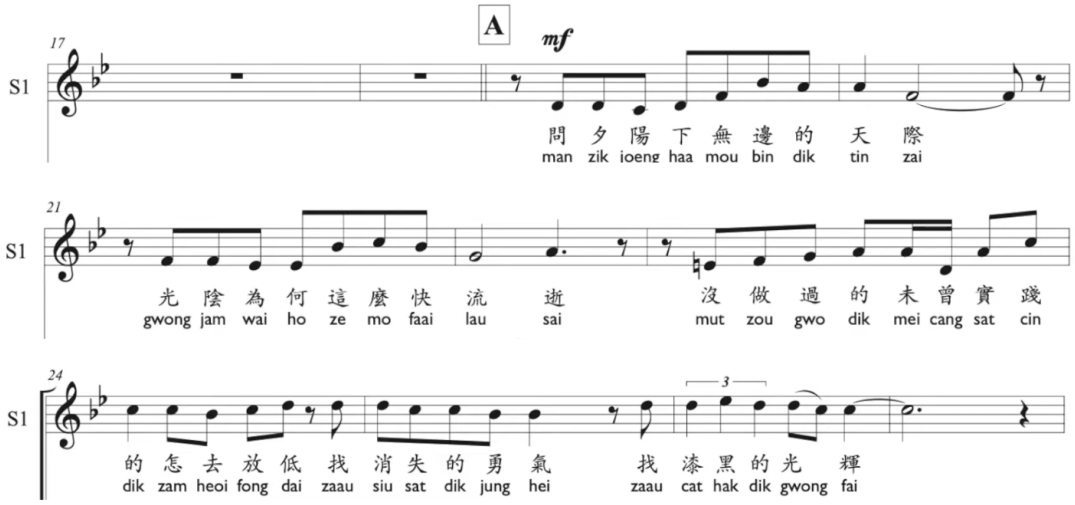


Figure 7.7 "A New Day" (新的一天). Authorized by the composer.

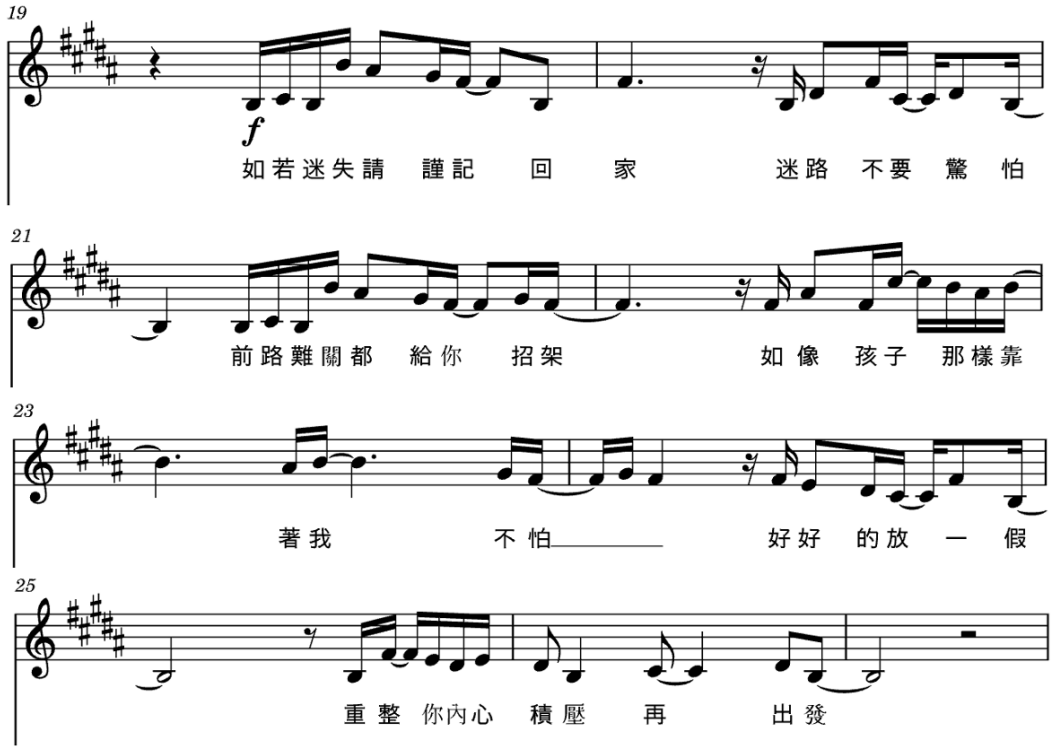


Figure 7.8 "Songs I Listen To When I Feel Lost" (当我迷失时会听的歌). Source: musescore.com

References

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047. https://doi.org/10.1126/science.aaz3468

Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., Du, Y., & Zatorre, R. J. (2024). Spectro-temporal acoustical markers differentiate speech from song across cultures. *Nature Communications*, *15*(1), 4835. https://doi.org/10.1038/s41467-024-49040-3

Aniruddh D. Patel. (2008). *Music, language, and the brain*. Oxford University Press. https://proxy.library.mcgill.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=209632&scope=site

Appelbaum, I. (1996). The lack of invariance problem and the goal of speech perception. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1541–1544. https://doi.org/10.1109/ICSLP.1996.607912

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*, *125*(2), 238–251. https://doi.org/10.1093/brain/awf028

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

Belfi, A. M., Karlan, B., & Tranel, D. (2016). Music evokes vivid autobiographical memories. *Memory*, *24*(7), 979–989. https://doi.org/10.1080/09658211.2015.1061012

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), Article 6767. https://doi.org/10.1038/35002078

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLOS ONE*, *8*(4), e60676. https://doi.org/10.1371/journal.pone.0060676

Brattico, E., Näätänen, R., & Tervaniemi, M. (2001). Context effects on pitch perception in musicians and nonmusicians: Evidence from event-related-potential recordings. *Music Perception*, *19*(2), 199–222. https://doi.org/10.1525/mp.2001.19.2.199

Burns, E. M., & Ward, W. D. (1978). Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *The Journal of the Acoustical Society of America*, *63*(2), 456–468. https://doi.org/10.1121/1.381737

Champely, S. (2020). *pwr: Basic functions for power analysis (R package version 1.3-0)* [Computer software]. https://CRAN.R-project.org/package=pwr

Chan, Marjorie, K. M. (1987). Tone and melody in Cantonese. Berkeley Linguistic Society, *Proceedings of the Thirteenth Annual Meeting*. 26-37.

Chao, Y. R. (1956). *Tone, intonation, singsong, chanting, recitative, tonal composition, and atonal composition in Chinese*. https://books.google.ca/books?id=HnlbNQEACAAJ

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

Clarke, E. F. (1987). Categorical rhythm perception: An ecological perspective. *Action and perception in rhythm and music*, 55, 19-33.

Cummins, F. (2013). Joint speech: The missing link between speech and music? *Percepta - Revista de Cognição Musical*, *1*(1), Article 1.

de Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*(3), 463–506. https://doi.org/10.1111/j.1467-9922.2006.00374.x

Deutsch, D. (1986). A musical paradox. *Music Perception*, *3*(3), 275–280. https://doi.org/10.2307/40285337

Deutsch, D. (2013). Absolute pitch. In *The psychology of music, 3rd ed* (pp. 141–182). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-381460-9.00005-5

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, *129*(4), 2245–2252.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

Fant, G. (1971). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. Walter de Gruyter.

Feng, Y. (2022). *Categorical speech perception across the lifespan* [The Hong Kong Polytechnic University]. https://theses.lib.polyu.edu.hk/bitstream/200/11620/3/6149.pdf

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition*, *100*(1), 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal*

of the Acoustical Society of America, *119*(3), 1712–1726. https://doi.org/10.1121/1.2149768

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. https://doi.org/10.1016/S1364-6613(00)01839-8

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of Music*, *31*(3), 221–272. https://doi.org/10.1177/03057356030313002

Gandour, J. T. (1978). II - The Perception of Tone. In V. A. Fromkin (Ed.), *Tone* (pp. 41–76). Academic Press. https://doi.org/10.1016/B978-0-12-267350-4.50007-8

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Good, A., Gordon, K. A., Papsin, B. C., Nespoli, G., Hopyan, T., Peretz, I., & Russo, F. A. (2017). Benefits of music training for perception of emotional speech prosody in deaf children with cochlear implants. *Ear and Hearing*, *38*(4), 455–464. https://doi.org/10.1097/AUD.0000000000000402

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006

Hallam, S. (2010). 21st century conceptions of musical ability. *Psychology of Music*, *38*(3), 308–330. https://doi.org/10.1177/0305735609351922

Harnad, S. (2003). *Categorical Perception*. Encyclopedia of Cognitive Science (01/12/03). Nature Publishing Group: Macmillan. https://eprints.soton.ac.uk/257719/

Herzog, G. (1934). Speech-melody and primitive music. *The Musical Quarterly*, *20*(4), 452–466.

Ho, Wing See Vincie. (2006). The tone-melody interface of popular songs written in tone languages. Paper presented at *the 9th international conference on music perception and cognition*, Bologna, August 22-26 2006.

Janata, P. (2009). The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex*, *19*(11), 2579–2594. https://doi.org/10.1093/cercor/bhp008

Jeffries, K. J., Fritz, J. B., & Braun, A. R. (2003). Words in melody: An H(2)15O PET study of brain activation during singing and speaking. *Neuroreport*, *14*(5), 749–754. https://doi.org/10.1097/00001756-200304150-00018

Johnson, K., & Sjerps, M. (2018). Speaker Normalization in Speech Perception. *UC Berkeley Phonology Lab Annual Reports*, *14*. https://doi.org/10.5070/P7141042474

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Kämpfe, J., Sedlmeier, P., & Renkewitz, F. (2011). The impact of background music on adult listeners: A meta-analysis. *Psychology of Music*, *39*(4), 424–448. https://doi.org/10.1177/0305735610376261

Kawahara, H., & Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana*, *36*(5), 713–727. https://doi.org/10.1007/s12046-011-0043-3

Kirby, J., & Ladd, D. R. (2016). Tone-melody correspondence in vietnamese popular song. *5th International Symposium on Tonal Aspects of Languages (TAL 2016)*, 48–51. https://doi.org/10.21437/TAL.2016-10

Koelsch, S. (2011). Toward a Neural Basis of Music Perception – A Review and Updated Model. *Frontiers in Psychology*, *2*. https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00110

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599–605. https://doi.org/10.1038/nrn2882

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PLOS ONE,* 7(12), Article e52508. https://doi.org/10.1371/journal.pone.0052508

Lenth R (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.11.2-00002*, https://rvlenth.github.io/emmeans/.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053–4062. https://doi.org/10.1121/1.3651816

Liberman, A. M. (1957). Some Results of Research on Speech Perception. *The Journal of the Acoustical Society of America*, *29*(1), 117–123. https://doi.org/10.1121/1.1908635

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Lin, H.-R., Kopiez, R., Müllensiefen, D., & Wolf, A. (2021). The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Musicae Scientiae*, *25*(2), 226–251. https://doi.org/10.1177/1029864919871987

List, G. (1961). Speech melody and song melody in central thailand. *Ethnomusicology*, *5*(1), 16–32. https://doi.org/10.2307/924305

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), Article 3139. https://doi.org/10.21105/joss.03139

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual Differences in Music Reward Experiences. *Music Perception*, *31*(2), 118–138. https://doi.org/10.1525/mp.2013.31.2.118

McMurray B. (2022). *The Myth of Categorical Perception*. PsyArXiv. https://doi.org/10.31234/osf.io/dq7ej

Mead, K. M. L., & Ball, L. J. (2007). Music tonality and context-dependent recall: The influence of key change and mood mediation. *European Journal of Cognitive Psychology*, *19*(1), 59–79. https://doi.org/10.1080/09541440600591999

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and function in human song. *Current Biology*, *28*(3), 356-368.e5. https://doi.org/10.1016/j.cub.2017.12.042

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, *102*(3), 1864–1877. https://doi.org/10.1121/1.420092

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: Association with lexical tone agnosia. *Brain*, *133*(9), 2635–2642. https://doi.org/10.1093/brain/awq178

NIRx Medical Technologies, LLC. (2021). *NIRx Arora* [Computer software]. Retrieved from https://nirx.net/arora

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J. M., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., Hegde, S., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D. P., Sadaphal, P. M., Hadavi, S., … Savage, P. E. (2024). Globally, songs and instrumental melodies are slower and higher and use more stable pitches than speech: A registered report. *Science Advances*, *10*(20), eadm9797. https://doi.org/10.1126/sciadv.adm9797

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681. https://doi.org/10.1038/nn1082

Patel, A. D. (2011). Language, music, and the brain: A resource-sharing framework. In *Language and Music as Cognitive Systems*. https://doi.org/10.1093/acprof:oso/9780199553426.001.0001

Patel, A. D., Foxton, J. M., & Griffiths, T. D. (2005). Musically tone-deaf individuals have difficulty discriminating intonation contours extracted from speech. *Brain and Cognition*, *59*(3), 310–313. https://doi.org/10.1016/j.bandc.2004.10.003

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. https://doi.org/10.1016/j.wocn.2010.09.003

Peretz, I. (2016). Neurobiology of Congenital Amusia. *Trends in Cognitive Sciences*, *20*(11), 857–867. https://doi.org/10.1016/j.tics.2016.09.002

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of Musical Disorders. *Annals of the New York Academy of Sciences*, *999*(1), 58–75. https://doi.org/10.1196/annals.1284.006

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*(7), Article 7. https://doi.org/10.1038/nn1083

Pierce, J. R. (1992). *The science of musical sound* (Rev. ed). W.H. Freeman. https://cir.nii.ac.jp/crid/1971712334742210594

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition*, *35*(2), 242–253. https://doi.org/10.3758/BF03193445

Rakowski, A. (1990). Intonation variants of musical intervals in isolation and in musical contexts. *Psychology of Music*, *18*(1), 60–72. https://doi.org/10.1177/0305735690181005

Reese, H. W. (2013). *The perception of stimulus relations: Discrimination learning and transposition*. Academic Press.

Repp, B. H. (1984). Categorical Perception: Issues, Methods, Findings. In *Speech and Language* (Vol. 10, pp. 243–335). Elsevier. https://doi.org/10.1016/B978-0-12-608610-2.50012-1

Rolls, E. T., Joliot, M., & Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122, 1-5. https://doi.org/10.1016/j.neuroimage.2015.07.075

Schellenberg, M., & Gick, B. (2018). Microtonal variation in sung cantonese. Phonetica, 77(2), 83–106. https://doi.org/10.1159/000493755

Schellenberg, M. H. (2013). *The Realization of Tone in Singing in Cantonese and Mandarin*. 168.

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J.-L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *NeuroImage*, *51*(1), 450–461. https://doi.org/10.1016/j.neuroimage.2010.02.023

Schulze, H.-H. (1989). Categorical perception of rhythmic patterns. *Psychological Research*, *51*(1), 10–15. https://doi.org/10.1007/BF00309270

Serafine, M. (1984). Integration of melody and text in memory for songs. *Cognition*, *16*(3), 285–303. https://doi.org/10.1016/0010-0277(84)90031-3

Shepard, R. N. (1964). Circularity in judgments of relative pitch. *The Journal of the Acoustical Society of America*, *36*(12), 2346–2353. https://doi.org/10.1121/1.1919362

Siegel, J. A., & Siegel, W. (1977). Categorical perception of tonal intervais: Musicians can't tellsharp fromflat. *Perception & Psychophysics*, *21*(5), 399–407. https://doi.org/10.3758/BF03199493

Stock, J. P. J. (1999). A reassessment of the relationship between text, speech tone, melody, and aria structure in beijing opera. *Journal of Musicological Research*, *18*(3), 183–206. https://doi.org/10.1080/01411899908574757

Tao, R., Zhang, K., & Peng, G. (2021). Music Does Not Facilitate Lexical Tone Normalization: A Speech-Specific Perceptual Process. *Frontiers in Psychology*, *12*, 717110. https://doi.org/10.3389/fpsyg.2021.717110

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences*, *109*(46), 19027–19032. https://doi.org/10.1073/pnas.1210344109

Tierney, A., & Kraus, N. (2013). Music training for the development of reading skills. *Progress in Brain Research*, *207*, 209–241.

Trainor, L. J., & Trehub, S. E. (1993). Musical context effects in infants and adults: Key distance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(3), 615–626. https://doi.org/10.1037/0096-1523.19.3.615

Vanden Bosch der Nederlanden, C. M., Qi, X., Sequeira, S., Seth, P., Grahn, J. A., Joanisse, M. F., & Hannon, E. E. (2023). Developmental changes in the categorization of speech and song. *Developmental Science*, *26*(5), e13346. https://doi.org/10.1111/desc.13346

Wallace, W. T. (1994). Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1471–1485. https://doi.org/10.1037/0278-7393.20.6.1471

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception & Psychophysics*, *64*(2), 198–207. https://doi.org/10.3758/BF03195786

Willems, R. M., & Peelen, M. V. (2021). How context changes the neural basis of perception and language. *iScience*, *24*(5), 102392. https://doi.org/10.1016/j.isci.2021.102392

Wong, P. C. M. (1998). *Speaker normalization in the perception of Cantonese level tones*, MS thesis, University of Texas at Austin, Austin, TX.

Wong, P. C. M., & Diehl, R. L. (2002). How Can the Lyrics of a Song in a Tone Language Be Understood? *Psychology of Music*, *30*(2), 202–209. https://doi.org/10.1177/0305735602302006

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. https://doi.org/10.1044/1092-4388(2003/034)

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063–1074. https://doi.org/10.1121/1.2213572

Yip, M. (2002). *Tone*. Cambridge University Press.

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

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547–558. https://doi.org/10.1038/nrn2152

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience*, *14*(4), 1908–1919. https://doi.org/10.1523/JNEUROSCI.14-04-01908.1994

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science (New York, N.Y.)*, *256*(5058), 846–849. https://doi.org/10.1126/science.1589767

Zatorre, R. J., & Gandour, J. T. (2007). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33–45. https://doi.org/10.1111/nyas.13410

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1252–1268. https://doi.org/10.1037/xhp0000216

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & Wang, W. S.-Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*, *124*, 536–549. https://doi.org/10.1016/j.neuroimage.2015.08.064

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

Zimeo Morais, G. A., Balardin, J. B., & Sato, J. R. (2018). fNIRS Optodes' Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports*, 8(1), 3341. https://doi.org/10.1038/s41598-018-21716-z