The phonology of tone and intonation in Guanzhong Mandarin

Jiarui Zhang

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2025

To my parents.

Acknowledgements

I would like to sincerely thank my supervisors, Professor Aditi Lahiri and Dr. Holly Kennard, for their guidance and support throughout my doctoral journey at Oxford. Without their encouragement, I would not have been able to complete this work. They taught me to consider the broader picture while attending to finer details, a perspective that I will carry forward. I am also deeply grateful to Professor Henning Reetz, Dr. Isabella Fritz, and Dr. John Lowe; their help made this experience even more rewarding.

Many thanks to the friends I met at Balliol College: Javier Navarro, Hrushikesh Loya, Aston Saini, Sophie Forst, Kristine Guillaume, Robert Ewart, Marcel Roed, Xinran Zhao, Titiksha Mohanty, and Lilja Saeboe. I am also grateful to Akash Malhotra, an old friend of many years. Their wit, intellectual conversations, and humour made even the dullest days, particularly during the pandemic, far more enjoyable. I truly appreciate their friendship and the laughter they brought into my life.

I am also thankful for the comments and advice I received at conferences, including Speech Prosody, the Linguistic Society of America Annual Meeting, the Old-World Conference in Phonology, and LabPhon. I am very grateful to my examiners, Professor Wolfgang de Melo and Dr. Cong Zhang, whose studies on tone and tune in Tianjin Mandarin led to in-depth discussion in this thesis. I thank them for their time in reading my work and for their advice.

Last, but by no means least, a huge thank you to my parents, Yanqin and Bin, whose unconditional love has been the foundation of all my academic pursuits and personal growth,

and to my husband, Thomas, for his support and love. I must also acknowledge my cherished companions, Bob and Kevin, who have brought me joy throughout this journey.

Abstract

In English, a rising intonation can turn a word into a question. Offering someone a coffee, for instance, involves simply saying “coffee?” with a rising pitch. Such intonational tunes carry both phonological properties (e.g. an LH rising contour) and semantic functions (e.g. offering). In tonal languages, however, where pitch already serves a lexical function, it remains unclear how intonational tunes are planned and represented, and how they interact with lexical tones.

This thesis investigates the status of intonational tunes in Guanzhong Mandarin (GuanM), a northern Chinese dialect with four lexical tones, and their interaction with lexical tone through two production studies, a perception study, and a psycholinguistic study.

The production studies, which used disyllabic words both in isolation and within longer utterances under focus, establish that syntactically unmarked yes-no questions in GuanM are characterised by a global high pitch register and a high boundary tone aligned with the right edge of the intonational phrase. This boundary tone is phonologically rising but results in different phonetic realisations of the phrase-final syllable: it enhances rising contours, converts level tones into rising contours, and prevents falling tones from completing their downward trajectories. Crucially, lexical tonal categories remain intact under question intonation even after lexical tonal alternation, indicating that intonation does not override lexical tone but is instead superimposed on the tonal system, affecting the phonetic realisation of tones without compromising their underlying phonological identity. Statements, by contrast, display a global low pitch register with declination. Focus is primarily marked by greater intensity and a faster speech rate, rather than by pitch.

This superimposition also has consequences for perception. Using the AX discrimination paradigm, the perception experiment shows increased perceptual difficulty and reduced discrimination sensitivity when lexical tone and intonational tune share similar pitch trajectories, particularly when both exhibit rising contours.

Finally, the psycholinguistic experiment returns to how tunes are planned and represented. Using the picture-word interference paradigm, it suggests that intonational tunes are accessed during early conceptual planning, setting up prosodic expectations that shape subsequent lexical retrieval. Different lexical tones interact differently with the interrogative tune, with retrieval facilitated when a tone's contour aligns naturally with the intonational structure.

Together, these findings establish the intonational system of GuanM and show that, in tonal languages, intonational tunes function as abstract prosodic templates: activated early in speech planning and interacting with lexical tones throughout production.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction and Background

1.1 Classifications of Chinese Languages

China has a vast territory and is rich in languages and dialects, with over 200 dialects spoken by millions of people across the country. Languages and dialects are reflections of society, history, and identity. The study of these dialects provides opportunities to investigate fundamental questions of linguistics, particularly in phonology and phonetics, such as tone and intonation, both in terms of language-specific features and universal linguistic principles. This thesis focuses on the prosodic systems, particularly the interaction between tone and intonation, of Guanzhong Mandarin (hereafter, GuanM), a variety of Mandarin, with broader implications for phonological theory.

Chinese dialects are categorised predominantly into either seven or ten groups (the latter classification has gained popularity in recent years). Criteria for classification are based on diachronic changes, including the phonological systems of most Chinese dialects that have existed since the Middle Chinese period[1], and on synchronic aspects, including differences between modern dialects and their specific contemporary phonological and lexical features (Kurpaska, 2010). The seven major dialect groups are Mandarin, Wu, Xiang, Gan, Hakka, Min, and Yue, as documented in Li and Thompson (1981) and Ramsey (1987). However, since the publication of the Language Atlas of China (Wurm et al., 1987), the classification into ten groups including the three new groups, Jin, Pinghua, and Hui, has gained support from state officials. This ten-group classification provides a comprehensive framework for understanding linguistic diversity in China, while recognising both the preliminary nature of dialect classification and the existence of varieties that remain unclassified, such as Shaozhou patois

[1] Min split off already in the Old Chinese period.

and Waxiang. It also reflects the historical processes that have shaped the development of Sinitic languages (Chappell, 2001). The geographic distribution of these ten dialect groups is illustrated in Figure 1.1.

Figure 1.1 Map of Chinese dialect groups (adapted from Wyunhe, 2011)

Mandarin can be further divided into eight subdivisions: Northeastern, Beijing, Jilu, Jiaoliao, Central Plains, Lanyin, Jiang-Huai, and Southwestern Mandarin (Wurm et al., 1987). Of these, Putonghua "common language", or Standard Chinese (SC), was established based on the northern dialects, particularly the Mandarin dialects, with Beijing pronunciation as its phonetic basis (Kurpaska, 2010). The terms SC, (Standard) Mandarin, and Putonghua are used

interchangeably throughout this study. SC is the most prominent variety within the Mandarin family and shares a common logographic system with other Mandarin varieties while maintaining substantial lexical and syntactic similarities.

1.2 Shaanxi Province and its Dialects

In Figure 1.1 above, the circled area highlights the Xi'an region of Shaanxi Province where GuanM is predominantly spoken. Xi'an, the capital city of Shaanxi Province located in northwestern China, has played a critical role in Chinese history as the starting point of the ancient Silk Road and the home to 13 Chinese dynasties. With a population of over 39 million in Shaanxi Province[2], of whom approximately 13 million reside in Xi'an, the region has a variety of dialects. The Jin dialect dominates the northern prefectures, including 19 towns (Xing, 2007), while Central Plains Mandarin, to which GuanM belongs, predominates in the southern prefectures. The southern prefectures also contain small populations of Southwestern Mandarin and Jiang-Huai Mandarin speakers. The term GuanM, also known as the Xi'an or Guanzhong dialect, has both a narrow and a broad interpretation. Narrowly, it refers to the variety spoken in Xi'an itself; more broadly, it encompasses varieties spoken across the surrounding region, including Gaoling, Lantian, Weinan, and Baishui (Xing, 2007). Furthermore, GuanM has two distinct sub-dialects, which are the Xifu dialect, spoken in western prefectures such as Baoji, and the Dongfu dialect, predominating in the eastern prefectures including Xi'an, Xianyang, and Weinan. The latter, the Dongfu dialect, widely regarded as the representative of GuanM, is the primary focus of the present investigation. Figure 1.2 presents the distribution of dialects within Shaanxi Province based on Xing's (2007)

[2] The data were sourced from the official website of the Shaanxi government (https://www.shaanxi.gov.cn/).

descriptions, highlighting the distinctions between the Dongfu and Xifu dialects, as well as among other dialect varieties found in the province.

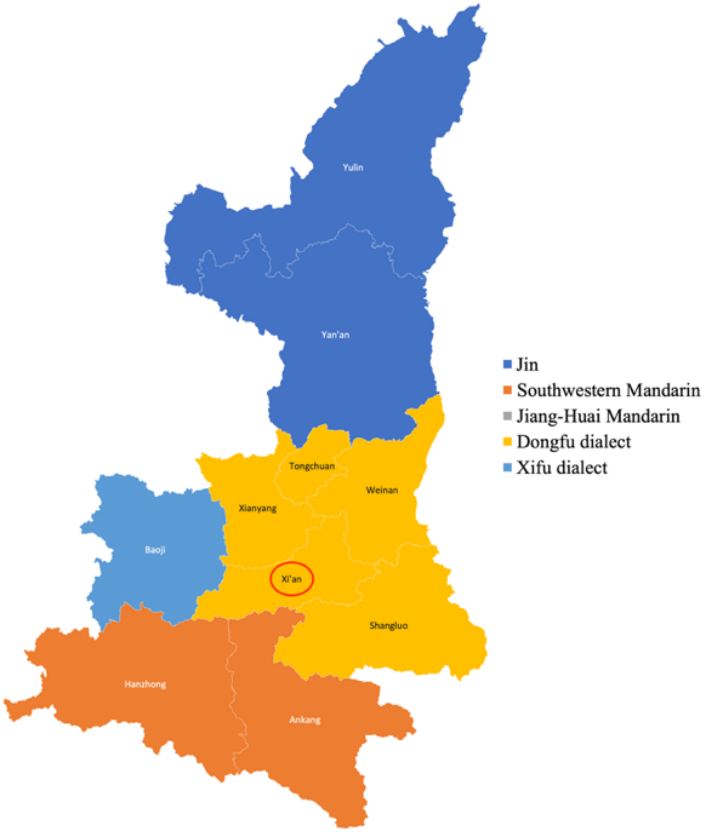


Figure 1.2 Map of the distribution of Shaanxi dialects

1.3 Attitudes towards SC and GuanM

Since the establishment of the People's Republic of China (PRC) in 1949, the standardisation and promotion of a unified Chinese language have been key aspects of China's language policy. The Central Committee for Language Reform, established in the same year as the PRC, proposed the adoption of Northern Chinese dialects as the basis for a standardised national language, driven by the need for cohesive linguistic standards to improve national literacy and communication.

Over the decades, the promotion of SC has progressed through three critical stages that firmly established SC (Putonghua) as the national common language across education, administration, media, and public communication. The first milestone was the official recognition of Putonghua as the standard form of Chinese in 1956 by the State Council in Promoting Putonghua's guidelines, which marked the beginning of institutionalised language standardisation. The second phase started in 1982 when the Constitution recognised Putonghua as a legal requirement under Article 19, leading to extensive language standardisation mandates in educational laws and regional autonomy statutes. Government media also adopted Putonghua as the official broadcast language. The third phase commenced in 2000 with the Law on Standard Spoken and Written Chinese Language, which consolidated the status of Putonghua and established detailed implementation guidelines across education, media, and public administration. Local governments subsequently enacted corresponding regional regulations (J. Yang, 2019). By the end of 2021, the PRC's Ministry of Education set a goal to have 85% of China's population speak Putonghua fluently by 2025. On June 28, 2022, the Ministry announced significant progress, as 95% of the educated population now uses standardised Chinese characters, and the illiteracy rate has dropped to 2.67%[3].

As SC becomes more widespread in daily life, it is common for people to alternate between SC and their local dialect. This pattern also applies to residents of Shaanxi Province, who engage in daily code-switching between SC and GuanM. H. Zhang et al. (2017) discovered that SC prevails in formal and professional settings, with just 3.19% of participants using only GuanM in the workplace and 37.93% using both varieties. The use of SC mainly appears among younger generations who often view GuanM as obsolete or unsophisticated. Moreover,

[3]The data were sourced from the official website of the Chinese government (https://www.gov.cn/).

only 42.03% of participants fully understand GuanM, mainly those born between the 1950s and 1970s, while millennial generations show considerably less interest or proficiency in GuanM. Ren and Yang (2012) investigated attitudes and usage patterns of SC and GuanM in Xi'an. Their findings showed that SC, carrying overt prestige, is regarded as the main pathway for career and educational advancement. GuanM, by contrast, holds covert prestige and thrives in informal settings like family conversations and local markets. This indicates a stable diglossia in which both varieties serve distinct yet complementary social functions. Attitudes towards both remain largely positive, with only 20% of participants reporting they use SC to signal social dominance. The persistence of GuanM demonstrates the relationship between the historical identity of Xi'an and present-day sociolinguistic pressures, as speakers engage in ongoing negotiation between dialect maintenance and adaptation to SC in institutional settings, particularly given the widespread bilingualism among speakers of GuanM and SC.

My personal language background has motivated this research. I spoke GuanM during my early years, but my schooling and social interactions were mainly in SC. Eventually, this resulted in a gradual shift: although my understanding of GuanM is still native-like, my active usage has decreased dramatically. My personal journey of navigating between SC and GuanM, along with observing younger generations show declining use of local varieties, has fuelled my curiosity about exploring prosody in GuanM.

1.4 Research Goals of the Present Study

Chinese languages and dialects use lexical tones to convey different meanings, which sets them apart from non-tonal languages. For example, the syllable [ma] in SC can mean 'mother' with T1 (high-level), 'numb' with T2 (rising), 'horse' with T3 (dipping), or 'scold' with T4 (falling). The tone is realised through variations in F0, the acoustic correlate of pitch. The F0 pattern

serves as the foundation for intonational patterns as well, which encode sentence-level meanings like focus, emotion, or modality. In tonal languages like Mandarin, F0 serves a dual function, conveying both lexical and post-lexical prosodic information simultaneously. Separating these two components can be difficult because it requires identifying the acoustic features that correspond to tones, including pitch height, contour, and duration at the syllable level, as well as identifying features that signal intonation, such as pitch range and boundary tones at the phrasal level.

This thesis examines whether AM Theory, a widely used model for intonation, applies to tonal languages like GuanM. The AM framework was originally developed for non-tonal languages, by modelling intonation through pitch targets and hierarchical structures. Given the dual functions of F0 in tone and tune in Chinese languages, a key question arises: Can AM Theory capture the tone-tune interaction in GuanM, or does it need refinements to accommodate the prosodic features of the dialect?

To address this question, this thesis investigates the syntactically unmarked yes-no question intonation in GuanM, focusing on tone-tune interaction and addressing gaps in current phonological and prosodic studies of GuanM, since previous work has mainly relied on impressionistic descriptions or historical documentation of tonal phenomena. This research aims to formalise the analysis by incorporating both theoretical and empirical approaches. Production studies draw on the AM framework; at the same time, perceptual and psycholinguistic experiments offer complementary perspectives of tone-tune interactions from different methodological angles.

Therefore, this thesis has two aims: (1) to formalise the prosodic study of GuanM through theoretical frameworks and practical approaches, and (2) to enhance the overall understanding of tone-tune interactions in tonal languages. By combining these research strands, the study seeks to deepen our understanding of the prosodic system of GuanM while proposing broader implications for linguistic theory. The following section outlines the thesis structure, with section 1.6 then introducing GuanM in detail.

1.5 Structure of the Thesis

Chapter 2 provides the theoretical foundation for this thesis, Autosegmental-Metrical (AM) Theory. It introduces the various intonation models and explains why the AM framework is adopted. It also explores the development of the theory, tenets, phonetic realisation, and application to Chinese languages.

Chapters 3 and 4 present production studies, examining the tone-tune interaction in GuanM. Chapter 3 starts the empirical investigation by analysing the production of questions using isolated disyllabic words, with the focus on tone-tune interaction in tonal alternation contexts. The chapter includes a detailed acoustic examination by assessing F0 parameters such as F0 mean, maximum, minimum, and range. It also measures intensity, duration, and the timing of F0 peaks and valleys. These measurements help us understand how intonation and tone work in isolated contexts, which establishes the groundwork for more complex scenarios in later chapters. Chapter 4 examines disyllabic words in the focus position within longer utterances. It investigates the complex interactions among tone, tune, and focus using the same acoustic measurements as Chapter 3.

Chapter 5 moves to perception, using an AX paradigm experiment to explore how listeners process and interpret the relationship between tone and tune. The study examines how listeners perceive question tunes while concurrently processing the tone-tune interaction, a task that requires both storing and accessing tune details alongside lexical tones. It also investigates whether listeners depend on particular acoustic cues such as pitch height, pitch contour, or duration to distinguish the question tune.

Chapter 6 adopts a psycholinguistic approach to explore how the question tunes are planned in relation to tone, using the picture-word interference paradigm. The study investigates the processes involved in real-time speech production of tune, showing how speakers organise and implement tunes while managing tone requirements. It also explores how semantic and phonological auditory distractors during the picture-naming task affect the planning and execution of tunes.

Chapter 7 brings together the results from all studies. It assesses whether the research goals have been achieved and also explores wider implications for tone and tune interaction.

1.6 Guanzhong Mandarin

1.6.1 Consonants and Vowels

This section presents the segmental inventory, including vowels, consonants, and syllable structure in GuanM.

GuanM has 23 consonants, as shown in Table 1.1[4].

[4] In the table, vertical positioning indicates aspiration (unaspirated in the first line and aspirated in the second line), while horizontal positioning represents voicing (voiceless on the left and voiced on the right).

Table 1.1 Consonants in GuanM

	Bilabial	Labiodental	Dental	Alveolar	Retroflex	Alveolo-palatal	Velar
Plosive	p p^h			t t^h			k k^h
Nasal	m			n			ŋ
Fricative		f v		s	ʂ ʐ	ɕ	x
Affricate				ts ts^h	ʈʂ $ʈʂ^h$	tɕ $tɕ^h$	
Lateral approximant				l			

Some scholars propose a larger inventory of 25 consonants (e.g., Lan, 2011; Sun, 2007; Wang & Li, 1996; W. Zhang, 2002) or 26 consonants (e.g., Beijing Daxue Zhongguo Yuyan Wenxuexi Yuyanxue Jiaoyanshi, 2003). These discrepancies arise from treating context-dependent allophones as distinct phonemes, most evidently in two cases. The first concerns the nasals: the alveolar /n/ acts as the underlying phoneme, surfacing as [ȵ] only when palatalised by high front vowels (e.g., [ȵy52] '女, woman'; [ȵi24] '泥, mud'), while [n] appears elsewhere (e.g., [nan24] '南, south'). In Western GuanM and Southwestern Shaanxi dialects, the initial [n] has shifted to [l] when preceding syllables with the vowels [a], [o], and [u], a feature that distinguishes these varieties from SC (Sun & Fu, 2019). For instance, words include [lan24] '男, man' (SC: [nan35]), [lu24] '奴, slave' (SC: [nu35]), and [lau55] '恼, mess' (SC: [nɑʊ51]).

Second, the labiodental affricates [pf] and [pf^h], segments absent in SC, are allophonic variations of the retroflex affricates /ʈʂ/ and /$ʈʂ^h$/. /ʈʂ/ and /$ʈʂ^h$/ undergo labiodentalisation when preceding the rounded vowel [u] or diphthongs and triphthongs beginning with this vowel (Qi, 2011). Examples include words such as [pfu31] '猪, pig' and [pf^hu31] '出, out'. In all other

environments, they surface as retroflexes [ʈʂ] and [ʈʂʰ] (e.g., [ʈʂau31] ‘招, trick’ and [ʈʂʰɤ31] ‘车, car’). By recognising these complementary distributions of consonants, this study maintains the consonant count at 23.

A further distinction between GuanM and SC is the presence of the voiced labiodental fricative [v], which occurs syllable-initially in GuanM but is absent in SC. For instance, SC typically realises words such as 袜 ‘sock’ and 雾 ‘fog’ as onsetless syllables [ua51] and [u51], respectively. GuanM speakers in regions like Qingjian, Zichang, and Yanchuan (Northern Shaanxi) pronounce them with the [v] onset as [va31] and [vu55], respectively[5].

As observed, there is no one-to-one correspondence between consonants in GuanM and SC, although the two varieties share the same consonants in the majority of lexical items. These differences can be traced back to the common ancestor, Middle Chinese. For example, both varieties maintain a phonemic contrast between alveolar sibilants [ts, tsʰ, s] and retroflex sibilants [ʈʂ, ʈʂʰ, ʂ], but in present-day GuanM, there are fewer characters realised with retroflex sibilants than in SC. Specifically, syllables whose onsets derive from the Middle Chinese retroflex sibilant series (庄组 Zhuangzu, Divisions II and III), the palatal sibilant series (章组 Zhangzu, 止摄 Zhi rhyme class, Division III), and the retroflex stop series (知组 Zhizu, Division II) are nowadays pronounced with alveolar sibilants in non-labialised (开口呼

[5] At present, the realisation of onset [v] shows regional variation within GuanM. According to Sun and Fu (2019), while most areas in Northern Shaanxi (including the aforementioned regions) alongside Baoji, Shangzhou, Luonan, and Danfeng retain the [v] onset, other areas such as Linyou, Xunyi, and the majority of Southern Shaanxi realise these syllables with the vowel [u].

Kaikouhu) syllables[6], whereas the corresponding syllables are realised with retroflex sibilants in SC (Sun, 1997). An example of this comparison between SC and GuanM is given below.

Syllable	渣	柴	渗
	'residue'	'firewood'	'seep'
SC	[tʂa55]	[tʂʰai35]	[ʂən51]
GuanM	[tsa31]	[tsʰæ24]	[sen55]

The vowel inventory of GuanM consists of eight monophthongs (as illustrated in Figure 1.3), eleven diphthongs, and three triphthongs. The diphthongs are [ia], [ie], [iæ], [ye], [yo], [ua], [uæ], [uo], [ei], [ou], and [au], while the triphthongs are [iau], [iou], and [uei].

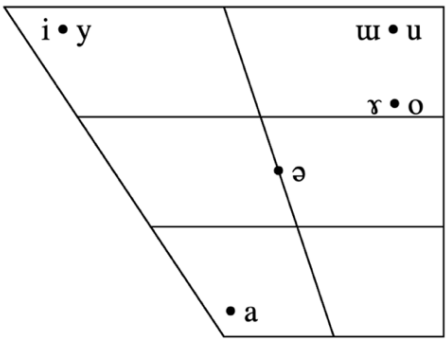


Figure 1.3 Vowels in GuanM

1.6.2 Syllable Structure

The syllable structure of GuanM is similar to SC. Both follow a maximum 'CGVX' pattern. In this template, C is an optional consonant. G is a glide ([j], [w]). V is a vowel. X is a nasal coda, limited to [n] or [ŋ]. Some scholars (e.g., Cheng, 1966; Howie, 1976; Zee & Lee, 2001) view the pre-nuclear G component as a vowel (i.e., /i/, /u/, /y/, /ɯ/) rather than a glide.

[6] Middle Chinese has 36 initial consonants according to the traditional classification found in rhyme books, grouped by place and manner of articulation, which includes labials (e.g., 帮组 Bangzu), dentals (e.g., 端组 Duanzu), dental sibilants (精组 Jingzu), retroflex stops (知组 Zhizu), retroflex sibilants (庄组 Zhuangzu), palatal sibilants (章组 Zhangzu), velars (见组 Jianzu), laryngeals (影组 Yingzu), etc. The "divisions" (等 Deng) refer to a four-way classification of Middle Chinese syllables based on vowel quality and the presence of medial glides (see Baxter, 1992). '开口呼 Kaikouhu' syllables are pronounced with an open mouth, without the lip rounding of [u]/[y] or the tongue fronting of [i]/[y].

GuanM and SC differ in one important respect concerning the distribution of the velar nasal [ŋ]. In SC, [ŋ] cannot occur in onset position, whereas in GuanM it can function as an onset consonant, followed by vowels such as [a], [ɤ], and [ə]. This feature is preserved from Middle Chinese (Y. Wang, 1995). The distinction is exemplified in several words: 'goose' (鹅) is realised as onsetless [ɤ35] in SC but as [ŋo24] in GuanM; 'safe' (安) appears as [an55] in SC but as [ŋan31] in GuanM; and 'kindness' (恩) is pronounced as [ən55] in SC but as [ŋən31] in GuanM.

1.6.3 Tonology

1.6.3.1 Lexical Tones

Chao (1930) developed the five-point tone system, which is a numerical framework for representing pitch contours in tonal languages. The system divides the pitch range into four equal parts, creating five distinct levels: 1 (low), 2 (half-low), 3 (mid), 4 (half-high), and 5 (high), which captures lexical tones by using two or three numbers. As mentioned earlier, in SC, the syllable [ma] can convey entirely different meanings depending on its tone. With T1 (55), it means '妈, mother'. With T2 (35), it means '麻, numb'. With T3 (214), it means '马, horse'. With T4 (51), it means '骂, scold' (Chao, 1968).

GuanM also has four lexical tones, the same number as SC. However, researchers have suggested different tone values based on the 5-point scale. The value of T2 is mostly the same in all studies, but the values of the other tones vary. Table 1.2 displays the various tone values proposed by researchers on GuanM. Although there are some differences in values, the majority of researchers agree that the four tones in GuanM are characterised by these

fundamental pitch patterns: T1 is generally characterised by a low falling pitch; T2 represents an ascending tone; T3 is a high falling tone; and T4 is a high-level tone. This thesis adopts the tone values proposed by J. Y. Zhang and Shi (2009) and by Sun and Fu (2019): T1 = 31, T2 = 24, T3 = 52, and T4 = 55.

Table 1.2 Summary of tone values in GuanM (on the 5-point scale)

Studies	T1 T2 T3 T4
Gong (1995)	21 24 52 45
M. Ma (2005)	21 24 53 44
J. Y. Zhang & Shi (2009); Sun & Fu (2019); Y. Zhao (2017)	31 24 52 55
P. Li (2001)	31 24 42 55

Some impressionistic studies (e.g., P. Xu, 2006; Hui et al., 2020) have proposed specific correspondences between GuanM tones and SC tones. T1 corresponds impressionistically to a neutral-sounding tone. T2 in GuanM matches T2 in SC. T3 in GuanM aligns with T4 in SC. T4 in GuanM corresponds to T1 in SC. Although P. Li (2001) and J. Y. Zhang (2009) recognised such impressionistic observations, they added that T1 in GuanM corresponds to T3 in SC, but only if the ending pitch rises to a mid-high-level.

Liu et al. (2020) challenged the idea of one-to-one tone mapping between GuanM and SC, based on both production and perception data. Their findings showed that most tone pairs display differences in F0 contours, except for the level tone pair (T4 in GuanM and T1 in SC). To be specific, the GuanM rising tone (T2) displays a shallower F0 curve compared to the SC T2, despite having similar mean F0. The GuanM falling tone (T3) shows a lower F0 height compared to the SC T4. The GuanM low tone (T1) does not have the distinctive low-falling-

rising pattern of SC (T3). Tones in GuanM, except for the falling tone, are also generally shorter in duration than their SC counterparts, resulting in a more compact tonal space. Their perception study revealed that bidialectal speakers of SC and GuanM perceive most tone pairs (level, rising, and falling) as very similar or identical, despite their acoustic differences. The low contour pair was slightly more distinguishable perceptually. However, speakers still generally considered them similar.

As mentioned above, impressionistic observations have noticed the tonal correspondences between two varieties. This raises the question of whether the one-to-one mapping exists between their respective tone systems. To address this question, it is once again necessary to trace both varieties back to the common ancestor, Middle Chinese. Middle Chinese is traditionally analysed as having four tonal categories: Ping 平 'level', Shang 上 'rising', Qu 去 'departing', and Ru 入 'entering', which continue to serve as an important analytical framework in contemporary dialectological studies in Chinese languages. Within the four tonal categories, an additional division is observed between upper (Yin 阴) and lower (Yang 阳) registers, which are directly associated with the voicing of syllable onsets: voiceless onsets align with the Yin register, whereas voiced onsets correspond to the Yang register. Table 1.3 presents the tonal categories and their corresponding tone values in GuanM and SC, using Chao's tone number system.

Table 1.3 Tone categories and their tone values in GuanM and SC

Tone category	GuanM Tone value	SC Tone value
T1 阴平 Yinping	31 (low-falling)	55 (high-level)
T2 阳平 Yangping	24 (rising)	35 (rising)
T3 上声 Shangsheng	52 (high-falling)	214 (dipping)

T4 去声 Qusheng	55 (high-level)	51 (falling)

For the first three tonal categories in Middle Chinese (Ping 平 'level', Shang 上 'rising', Qu 去 'departing'), GuanM broadly follows the same pattern as SC in its development from Middle Chinese. Syllables with voiceless onsets in the Middle Chinese level tone become T1 (阴平), while those with voiced onsets become T2 (阳平). Syllables with fully voiced obstruent onsets in the Middle Chinese rising tone merge into T4 (去声), a process referred to as zhuo2shang3 gui1qu4 浊上归去 (Sun, 2007).

The main divergence between GuanM and SC lies in the treatment of the entering tone. In Middle Chinese, the entering tone consists of checked syllables ending in a final stop consonant [p], [t], or [k], which has been lost in most Mandarin dialects. As a result, these syllables have been redistributed among the remaining tones. In SC, entering tone syllables are scattered unpredictably across all four tones, making their distribution difficult to memorise. In contrast, the distribution in GuanM follows a more regular pattern: syllables with fully voiced obstruent onsets now belong to the rising tone (T2/阳平), while the remainder have overwhelmingly merged into the low-falling tone (T1/阴平), with only a small number of exceptions (Sun, 1997). Table 1.4 below illustrates the distribution of the Middle Chinese entering tone syllables in GuanM and SC with relevant examples, based on Sun (1997). The second row of the table provides the Middle Chinese transcriptions using the Baxter-Sagart Old Chinese reconstruction (Baxter & Sagart, 2014).

Table 1.4 Distribution of Middle Chinese entering tone syllables in GuanM and SC (with examples)

	Voiceless unaspirated						Voiceless aspirated				Voiced obstruent			Voiced sonorant		
Chinese character	八	福	笔	毕	茁	易	脱	铁	刻	洽	白	集	寂	玉	墨	幕
MC IPA	[peat]	[pjuwk]	[pit]	[pjit]	[tsrjwet]	[yek]	[t^hwat]	[t^het]	[k^hok]	[heap]	[baek]	[dzip]	[dzek]	[ngjowk]	[mok]	[mak]
Gloss	‘eight’	‘bless’	‘pen’	‘finish’	‘carve’	‘change’	‘peel off’	‘iron’	‘cut’	‘accord with’	‘white’	‘gather’	‘quiet’	‘jade’	‘ink’	‘curtain’
SC Tone	T1	T2	T3	T4	T2	T4	T1	T3	T4		T2		T4	T4		
GuanM Tone	T1				T2	T4	T1			T3	T2	T1		T1	T2	T4
SC IPA	[pa55]	[fu35]	[pi214]	[pi51]	[tʂuo35]	[i51]	[t^huo55]	[t^hie214]	[k^hə51]	[tɕ^{h}ia51]	[pai35]	[tɕi35]	[tɕi51]	[y51]	[mo51]	[mu51]
GuanM IPA	[pa31]	[fu31]	[pi31]	[pi31]	[pfuɤ24]	[i55]	[t^huɤ31]	[t^hiɛ31]	[k^hei31]	[tɕ^{h}ia52]	[pɛ24]	[tɕi31]	[tɕi31]	[y31]	[mei24]	[mu55]

1.6.3.2 Neutral Tone

As in other Mandarin dialects, the presence of the neutral tone (T0) in GuanM raises the fundamental question about whether it is truly toneless or retains an underlying tonal specification. In SC, Chao (1968) observed that the pitch level of T0 is largely determined by the preceding full-tone syllable. Based on impressionistic observations, he described T0 as half-low after T1, mid after T2, half-high after T3, and low after T4. This suggests that T0 is not completely toneless or inherently low, regardless of environment, but rather inherits its pitch properties from the preceding tone, an argument that has shaped much of the subsequent research on T0 realisation.

From a phonological perspective, Yip (2002) analysed T0 in SC using an autosegmental framework, which differentiates register (upper/lower) and tonal (H/L) tiers. According to Yip, T0 in SC is pre-specified for the register feature [-upper] but lacks an inherent tonal feature; consequently, its actual pitch height is determined by the preceding full tone. Specifically, T0 manifests as low [1] after T4 [51], mid [3] after T1 [55] and T2 [35], and mid-high [4] after T3 [214]. The following tone-spreading rules (1) illustrate this process. For the T3 sequence in particular, Yip (1980) argued that a special insertion rule applies: an H tone is inserted after T3 when no other tone follows it.

(1)

T1 + T0	[+upper] +	[-upper]	→	[+upper] +	[-upper]
	H ∧ H			H ∧ H	
T2 + T0	[+upper] +	[-upper]	→	[+upper] +	[-upper]
	L ∧ H			L ∧ H	
T3 + T0	[-upper] +	[-upper]	→	[-upper] +	[-upper]
	L ∧ L	H		L ∧ L	H
T4 + T0	[+upper] +	[-upper]	→	[+upper] +	[-upper]
	H ∧ L			H ∧ L	

As for T0 in GuanM, Zheng (2004) argues that it is essentially assimilated to T1, which means it lacks significant pitch distinction. By contrast, an alternative analysis treats T0 as an underlying low tone, which is phonetically realised as the absence of a distinct tone in weak syllables. Separately, Gong (1995) identifies three phonological environments in which T0 occurs.

The first environment involves the word-final position, affecting both grammatical particles and lexical suffixes. To be specific, toneless particles include [ɕjɛ31] '些, some', the tense marker [luɤ] '了', and the auxiliary [tei] '得'. Lexical suffixes include [tʰɤu24] '头, head' and [tsi52] '子, son'. These suffixes carry full tones when spoken in isolation. However, they lose their inherent values when serving as suffixes, which often assimilate to the preceding tone. Some examples are [ʂi24 tʰɤu3] '石头, stone', [li52 tʰɤu3] '里头, inside', [xaŋ55 tsi3] '巷子, alley', and [tan52 tsi3] '胆子, courage'.

The second environment involves reduplicated syllables. In these cases, the second syllable is neutralised to a mid pitch [3] (except when following T1), rather than keeping its base tone (Gong, 1995). Following T1, it is realised as low. Such a pattern can be found across lexical categories, including nouns such as [njaw52 njaw3] '鸟鸟, bird' and [wan52 wan3] '碗碗, bowl', and verbs such as [ʈʂʰaŋ24 ʈʂʰaŋ3] '尝尝, taste' and [kʰan55 kʰan3] '看看, see'. Furthermore, as in SC, this neutralisation can distinguish polysemes. The fully toned [tuŋ31 ɕi31] means '东西, east and west'; meanwhile, the neutralised form [tuŋ31 ɕi1] shifts the meaning to '东西, thing'.

Finally, T0 may emerge in specific tonal alternation contexts. In sequences of T3-T3 (HLHL) or T3-T4 (HLH), the second tone may be weakened into T0 (Zheng, 2004), which is observed in words such as [law52 fu52→2] '老鼠, rat' and [tɕjɤu52 tʰsæ55→2] '韭菜, chives'.

1.6.3.3 Tonal Alternation Rules

In GuanM, disyllabic sequences sharing the same tone are subject to two primary dissimilation rules (categorised as right-domain tonal alternations). In these sequences, the first tone undergoes dissimilation while the second tone remains unchanged. These rules are formally defined in (2) below:

(2) Dissimilation rules in GuanM[7]

a. T1 + T1 → T2 + T1: HL_{Lr} → LH (24) /__ HL_{Lr}

b. T3 + T3 → T1 + T3: HL_{Hr} → HL_{Lr} (31) /__ HL_{Hr}

Rule (2)a shows that, when two T1 tones occur consecutively, the first low falling T1 transforms into a rising T2, while the second T1 remains unaltered, as seen in examples such as [tʂʰəŋ31→24 ɕjen31] '中心, centre', [kuei31→24 tɕia31] '国家, nation', [faŋ31→24 fa31] '方法, method', [tɕi31→24 liɛ31] '激烈, intense', and [kʰæ31→24 xua31] '开花, blossom'. However, Gong (1995) observes that this dissimilation is blocked if the second tone is reduced to a T0. This pattern can be found in words such as [kwəŋ31 tɕi31→1] '公鸡, rooster' and [tɕiæ31 tɕyɤ31→1] '街角, corner', where the first syllable remains [31] rather than changing to [24].

[7] Hr refers to the high register and Lr refers to the low register.

Rule (2)b applies to T3 + T3 sequences. In this environment, the first T3 changes to T1 (lowering its register while maintaining the contour shape), while the second T3 remains unchanged. Examples include [ʂɤu52→31 piau52] '手表, watch', [pau52→31 ɕian52] '保险, insurance', [ɕi52→31 ʂɤu52] '洗手, wash hands', and [fei52→31 kuɤ52] '水果, fruit'. This pattern appears to be a feature of an older variety of GuanM and is mainly attested among speakers born before 1990 (Gong, 1995; Yao, 2014).

1.7 Summary

The introductory chapter began with an overview of the linguistic landscape of China. It focused on the dialects of Shaanxi Province, where GuanM is primarily spoken. The chapter then outlined the research objectives and provided an overview of the thesis structure. This was followed by an introduction to GuanM, which included a description of its segmental and suprasegmental features. The next chapter discusses the theoretical background of the study.

Chapter 2 Autosegmental-Metrical Theory

This chapter presents the theoretical framework used in this thesis: AM Theory. The chapter is organised as follows. It begins with an overview of models in intonation research and places AM Theory within the broader context of intonational studies. The discussion then turns to the AM framework in greater detail, beginning with the historical development of AM Theory, followed by an explanation of its tenets and phonetic realisation. Finally, the chapter reviews the application of AM Theory to Chinese dialects.

2.1 Intonation Models

An effective intonation model should address two primary requirements. The first is the ability to explain variations in pitch contours observed across languages. The second is the capacity to generalise these contours, capturing how speakers perceive them as realisations of a shared underlying melody (Arvaniti, 2022). Arvaniti (2011) reviewed the major theoretical models proposed to account for intonation. These models differ primarily in three respects: whether they treat intonation contours as holistic units or as combinations of smaller elements, what they assume to be the basic phonological primitives of intonation, and how they relate intonational form to meaning.

At the most holistic level, configurational or gestalt models treat intonation contours as undivided whole units that directly reflect functional or structural aspects of speech. Early work, such as Bolinger (1951) and Jones (1976), argued that melodies should be viewed as continuous patterns that cannot be easily decomposed. English intonation, for example, is described in terms of a small number of basic tunes, such as rises and falls, which can be compressed or stretched to fit utterances of different lengths.

Modern versions of the gestalt approach include INTSINT (Hirst & Di Cristo, 1998; Hirst et al., 2000), which provides a symbolic transcription of entire F0 contours using relative pitch symbols such as Higher, Lower, Top, and Bottom; OXIGEN (Grabe et al., 2003), which utilises polynomials to model shape differences between sentence types; and PENTA (Y. Xu, 2005), which assigns pitch targets to every syllable while linking global contour shape directly to communicative functions such as statement versus question. However, Arvaniti (2011) argued that these models struggle to account for how a single melody varies systematically in shape across utterances of different lengths, as tunes do not simply stretch or shrink to fit the duration of the words.

Furthermore, the configurational tradition includes superpositional models, which suggest that intonation consists of a global trend and local perturbations that ride on that trend. For example, the IPO system (Hart et al., 1990) views localised movements as superposed on a larger declination component. Fujisaki's model (Fujisaki, 1983, 2004) suggests that global phrase commands generate overall F0 contours and local accent commands create pitch movements associated with prominent syllables at the same time. The models of Gårding (1983, 1987) and Thorsen (1980, 1985, 1986) link global pitch trends to utterance-level functions such as declarativity or interrogativity.

Unlike holistic approaches, level-based models use sustained pitch levels as their foundational units. Early level-tone models from American structuralism (Hockett, 1955; Pike, 1945; Trager & Smith, 1951) represented intonation using discrete pitch levels such as High, Mid, and Low, which were intended as abstractions similar to phonemes. However, these views were heavily criticised by Bolinger (1951) for failing to capture the continuous nature of pitch.

The most prominent current framework is AM Theory, which was developed by Bruce (1977) on Swedish and Pierrehumbert (1980) on English. Beckman and Pierrehumbert (1986) and Pierrehumbert and Beckman (1988) further advanced the theory. AM Theory represents intonation as an underspecified string of High and Low tones residing on an autosegmental tier. Unlike earlier approaches, this model focuses only on linguistically significant points, such as pitch accents on stressed syllables and boundary tones at the edges of phrases. By treating the pitch between these points as mere interpolation, AM Theory captures how the same melody remains recognisable even when the number of syllables in a sentence changes. A detailed description of this model is provided in the subsequent section.

There is an ongoing debate over the contrasting approaches to prosodic and intonational analysis proposed by AM and PENTA models. The PENTA (Parallel Encoding and Target Approximation) model, proposed by Y. Xu (2015), is an articulatory-based approach characterised as quasi-superpositional, articulatory and functional (Y. Xu, 2015). PENTA departs from traditional models by viewing prosody as the result of multiple communicative functions encoded in parallel. Each function impacts prosodic contours by making localised changes to pitch target parameters, such as height, slope, and approximation rate, operating through syllable-synchronised target approximation, in which articulatory movements attempt to achieve underlying pitch targets. With its functional orientation, PENTA prioritises the communicative roles of prosodic units over their auditory shapes, using language-specific encoding techniques to convert communicative intentions into articulatory parameters. A comparison of these two models is presented in the table below, based on Pierrehumbert's (2022) review.

Table 2.1 Comparison between the AM model and the PENTA model

Aspect	PENTA model	AM model
Key Concepts	• Configurational nature • Parallel encoding • Target approximation • Functional orientation	• Discrete tonal primitives • Association with segmental and prosodic structures • Underspecification • Compositionality • Cross-linguistic applicability
Levels of Representation	• Focuses on communicative functions and phonetic parameters	• Three levels: semantic, phonological, and phonetic
Structure of Representations	• Flat structures • Direct mapping from communicative functions to phonetic parameters	• Hierarchical representation of units (e.g., foot, segment, syllable, intonational phrase)
Phonological Function	• No intermediate phonological representation	• Intermediate phonological level acts as a bottleneck, reducing dimensionality and restricting mappings
Phonetic Implementation	• Sophisticated optimisation algorithms	• Simpler phonetic realisation algorithms
Main Advantages	• Superior phonetic implementation • Efficient optimisation • Better acoustic predictions	• Comprehensive linguistic framework • Handles complex interactions • Strong theoretical foundations
Main Limitations	• Less elaborate typological framework	• Complex to implement • Less accurate phonetic predictions

The strength of AM Theory lies in its ability to decompose intonational tunes (Arvaniti, 2022), since it can break down continuous pitch patterns into separate elements at the post-lexical level. The theory therefore creates abstract phonological representations, which emphasise similarities among different intonation patterns, allowing for a systematic examination of variation without focusing on specific phonetic details. These abstract forms are important in

the AM framework for three reasons. First, they predict how melodies align with speech segments. Second, they allow the same melodic patterns to apply to utterances of different lengths. Lastly, they facilitate the empirical analysis of actual speech patterns in different languages.

2.2 AM Theory

2.2.1 The Origin and Development

The term Autosegmental-Metrical (AM), introduced by Ladd (2008), refers to the integration of two subsystems of phonology required for analysing intonation: an autosegmental tier, representing the melodic aspects of intonation, and a metrical structure, accounting for prominence and phrasing. As the name suggests, AM Theory integrates both Autosegmental Theory and Metrical Theory, which are discussed below.

Segmental phonology dominated phonological research before the development of Autosegmental Theory. It analyses speech as a sequence of discrete units, or segments, which form the foundation of the sound system. Phonetic features and processes are used to associate and differentiate segments, as seen in generative phonology (Chomsky & Halle, 1968). Segments are ordered linearly, and features within each segment are unordered. Based on this feature-based framework, W. S.-Y. Wang (1967) proposed a systematic treatment of tone within generative phonology, arguing that tonal distinctions should be formally integrated into phonological representations, mirroring the treatment of segmental features. W. S.-Y. Wang's work is significant for applying the generative model to tonal languages, introducing a complete system of binary distinctive features for tone, including contour, high, central, mid, rising, and falling. Furthermore, his approach provides a rule-based account of tonal alternations, which is similar to the rules for stress and phonological transformations in The

Sound Pattern of English (Chomsky & Halle, 1968). Other scholars who explored tonal representation also treated tones as segmental features (e.g., Schachter & Fromkin, 1968; Woo, 1969).

Leben (1973) proposed the suprasegmental representation of tone, arguing that floating tones have no segmental realisation. His study of Mende nouns revealed five tonal melodies (L, H, HL, LH, and LHL), which can be distributed across one, two, or three syllables, challenging the constraints of the segmental framework. The constraints governing morpheme structure in Mende are independent of the number of vowels or syllables in a lexical entry. This independence therefore reinforces the need for a suprasegmental representation of tone.

Williams (1976) took this further. He claimed that at the deepest level of phonological representation, tones are not inherently linked to segments or syllables. Instead, they are linked to morphemes. Each morpheme contains an independent tonal string as part of its phonological representation. Thus, Williams proposed the Tone Mapping Rule, which relates the underlying string of tones in a phrase to its segmental (or syllabic) structure. Anderson (1976) also challenged the segmental framework by reconceptualising nasality as a suprasegmental feature, drawing parallels with tonal analysis. He argued that nasality in many languages behaves similarly to tones. Examples include prenasalisation and postnasalisation within single segments and nasal spreading across segment boundaries.

As mentioned above, traditional linear phonological models treat sounds as a single sequence, which struggle to account for phenomena such as tone spreading, deletion, or independence. Goldsmith (1976) developed Autosegmental Theory to address these limitations, built on earlier work by Leben (1973) and Williams (1976) that distinguishes tonal features from

consonants and vowels. Autosegmental Theory changes approaches to phonological analysis by introducing a nonlinear representation. Autosegments, such as suprasegmental features (e.g., tone), are represented on a separate tier, which is distinct from segmental elements like consonants and vowels. These tiers are linked through association lines, which indicate how features on one tier connect to elements on another, as illustrated in (3)a.

(3) a.

σ σ Segmental tier
| | Association line
T T Tone tier

b. Falling tone in SC

σ
|\
H L

A key feature of Autosegmental Theory is the autonomy and independence of autosegments from the segments they are associated with, which means that changes on the segmental tier, such as the deletion or insertion of vowels or consonants, do not necessarily affect tonal features, and vice versa. For example, in tonal languages, a tone may persist and reattach to a new segment even if its original vowel is deleted (e.g., tone persistence). Similarly, tones can shift without altering the consonant or vowel they are originally linked to. However, tones are only realised on the surface if they are connected to tone-bearing units (TBUs), such as syllables or moras. Furthermore, the theory treats contour tones (rising or falling pitches) as sequences of discrete level tones (e.g., High (H) and Low (L)). For example, a falling tone is represented as an HL sequence, as in (3)b. This representation can be used to explain phenomena such as tonal harmony in Yoruba, where a high tone spreads across multiple syllables, and tonal delinking, where a "floating tone" temporarily detaches before reattaching to another TBU.

Consequently, AM Theory captures essential properties of tones, including:

a. Mobility: A tone can shift from one segment to another.

b. Stability: A tone can persist even after the loss of its original segment, leading to processes including resyllabification and reassociation.
c. One-to-many association: A single tone may spread across multiple segments (tonal plateaux, tone spreading).
d. Many-to-one association: Multiple tonal features can be realised on a single segment (contour tones).
e. Toneless segments: Certain tone-bearing units do not inherently acquire phonological tones (Yip, 2002).

However, the association between tiers follows well-formedness conditions, supplemented by language-specific rules (Yip, 2002), which ensure a structured relationship between tones and TBUs, as outlined below:

(4) Well-formedness conditions (Yip, 2002)

a. Every TBU must have a tone.

b. Every tone must be associated to some TBU.

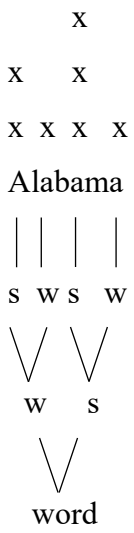
c. Association proceeds one-to-one, from left-to-right.

d. Association lines must not cross.

Metrical Theory, as proposed by Liberman (1975) and expanded by Liberman and Prince (1977), describes how stress works hierarchically in languages. The term metrical is derived from metrics, which studies how phonological units align within rhythmic structures. Phonological units can be smaller than words (for example, syllables and moras) or larger (for example, phonological phrases). This hierarchy shows how stress patterns work at different levels of language. Rather than treating stress as an isolated property of syllables, the theory represents relative prominence through two key structures: the metrical grid and the metrical

tree. As shown in (5), the metrical grid for the word Alabama, marked with "x" symbols, shows the alternation of strong (s) and weak (w) syllables, where column height reflects prominence. Below the word, the metrical tree maps how these strong and weak nodes relate to each other hierarchically.

(5)



The foundation of AM Theory was established by Pierrehumbert (1980) in her work on intonation in American English. She used a limited set of tonal elements, H and L tones, which are combined into structures known as pitch accents, phrase accents, and boundary tones, each fulfilling a specific role within the prosodic structure of an utterance. In Pierrehumbert's system, each intonation phrase must contain, at a minimum, a pitch accent, initial and final boundary tones, and a phrase accent. Therefore, her phonological inventory comprises monotonal pitch accents (H*, L*), bitonal pitch accents (H*+L, L*+H, H+L*, L+H*, H*+H), phrase accents (H-, L-), and boundary tones (H%, L%).

Later, Beckman and Pierrehumbert (1986) refined the model by introducing the concept of the intermediate phrase. They recognised that intonational phrasing operates on multiple levels, with smaller prosodic units embedded within larger ones. The intermediate phrase, marked by

a phrase accent, functions as a sub-constituent within the broader intonation phrase, which is typically delimited by boundary tones.

Ladd (1986) further developed the model by proposing a recursive model of intonational phrasing that allows for the coexistence of tonal and rhythmic cues at different levels of prosodic structure, accounting for cases where the tonal structure does not align neatly with rhythmic breaks or pauses, such as utterances containing multiple nuclear accents or intonational tags.

2.2.2 Tenets in AM Theory

Ladd (2008) has identified four tenets of AM Theory. The first tenet of AM Theory is the concept of sequential tonal structure, in which intonation is decomposed into a series of discrete tonal events including pitch accents and edge tones, which are phonologically specified at linguistically significant positions. In English, pitch accents are frequently linked to prominent syllables in the segmental string, and edge tones are associated with prosodic boundaries. Transitions between these tonal events are interpolated rather than explicitly represented in the phonological structure. For example, the rise-fall-rise contour is often used in strongly challenging or contradicting echo questions in English, as Ladd (2008) illustrates in (6):

(6) a. A: I hear Sue's taking a course to become a driving instructor.

B: Sue!?

b. A: I hear Sue's taking a course to become a driving instructor.

B: A driving instructor!?

In (6)a, a single-syllable expression, the intonation is interpreted as a blend of a rise-plus-fall pitch accent with a rising boundary tone, instead of being viewed as an entire intonation pattern. In (6)b, the contour similarly includes a rise-plus-fall pitch accent on driv- and a rising boundary tone, and the low-level pitch stretch on -ing instruct- serves as a transitional phase. Both have the same contour shape, differing only in the number of syllables, illustrating how intonational melodies adjust to various stress patterns and syllable counts.

The second tenet establishes the distinction between pitch accents and stress. Pitch accents are characterised as tonal features that mark intonational prominence, while stress refers to a structural property of syllables within a metrical hierarchy. Although pitch accents can signal prominence, stress may be conveyed through various phonetic cues, such as duration, intensity, or vowel quality. The distinction between pitch accent and stress can be illustrated by French, where the last full-vowel syllable in a French phonological word holds a structural prominence or "metrical strength" even in the absence of dynamic stress in the traditional sense. This syllable functions as an anchor for intonation tones, distinguishing itself structurally even when it is acoustically indistinct. By examining the tonal patterns in French sentences, specifically focusing on the behaviour of final schwas, Dell (1984) discovered that in cases where a final schwa is pronounced, intonation tones (such as high tones in emphatic statement contours) consistently avoid attaching to the schwa itself, rather targeting the preceding full-vowel syllable. This pattern is evident in phrases with or without right-dislocated elements. For example, *Il met la table, Mercier* 'He sets the table, Mercier' where the H tone remains aligned with the last strong syllable (ta- in table), rather than shifting to the schwa.

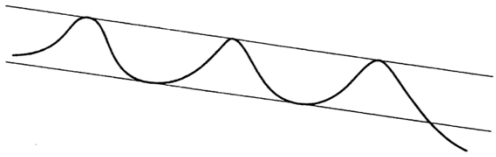
The third tenet emphasises the analysis of pitch accents in terms of level tones, specifically H and L tones in a bitonal system. AM Theory uses H and L tones as its basic tonal primitives,

analysing pitch accents and edge tones as combinations of these level tones, which therefore reduces the complexity of tonal inventories and maintains the ability to represent diverse intonational patterns at the same time. For example, in English, the four-level analysis (e.g., Trager & Smith, 1951) that describes six different falling contours (/21/, /31/, /41/, /32/, /42/, /43/) is unnecessary (Ladd, 2008). By representing all falling contours as HL and attributing changes in H or L height to external factors, AM Theory achieves a balance between phonological contrasts and paralinguistic variations.

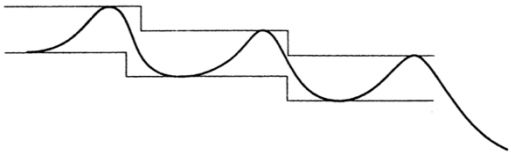
Finally, the fourth tenet of AM Theory accounts for global trends in pitch contours through local adjustments in the scaling of tones, such as the phenomenon of declination. This principle distinguishes between two approaches to describing an idealised downward-trending contour as shown in (7), adapted from Ladd (2008). The first approach (7)a conceptualises the contour as a result of global declination, where the pitch gradually lowers over the entire utterance. Many quantitative descriptions of declination treat it as a globally planned and encoded feature, representing an overarching contour that governs the intonational phrase. For example, in Fujisaki's (1983) model, the F0 contour is conceptualised as the superposition of a phrase component and one or more localised accent components, where the phrase component is represented by a continuous, gradually declining function that is generated independently of localised accentual events, thereby capturing the higher-level prosodic structure of the utterance. In the same way, the IPO model utilises a formulaic method to determine both the starting pitch level and the rate of decrease, with these factors affected by the utterance's duration (Hart et al., 1990). Therefore, declination is most accurately viewed as a top-down and holistic phenomenon that structures the pitch contour and indicates the end of an intonational phrase, as shown in (7)a.

However, AM Theory advocates a locally originating interpretation of declination, proposing that it arises naturally from the phonetic realisation of individual H and L tones within an utterance. Declination, from this perspective, is not globally encoded but is rather the cumulative result of successive local tonal adjustments, where each tonal target is scaled slightly lower than its predecessor, which has been observed in African languages, where the second H tone in an HLH sequence is realised phonetically lower than the first H tone, as shown in (7)b.

(7) a.



b.



In the cross-linguistic analysis of downtrends, it is essential to distinguish among declination, downstep, and downdrift. Declination is a phonetic phenomenon characterised by a gradual, time-dependent downward slope of F0. It is generally considered to lie outside phonological control (Downing & Rialland, 2016), and acts as a phonetic backdrop against which tones are scaled (Connell, 2001). Declination appears to be universal in declarative sentences but is often suspended in interrogatives. Downstep, on the other hand, is a phonological phenomenon that creates a distinct terracing effect (Connell, 2001). Once a tone is downstepped, all subsequent H tones remain at that lowered level. Stewart (1965) argued that the lowering of the second H in a HH sequence parallels the lowering found in HLH sequences. In the latter case (automatic downstep), the lowering results from the intervening L tone. In the former case (non-automatic

downstep), it is attributed to an underlying floating L tone, which has been lost historically but remains in the phonological structure. In some systems, non-automatic downstep is triggered by the Obligatory Contour Principle to separate adjacent H tones (Yip, 1988). Both types share the crucial terracing effect. The fact that downstep is phonemically contrastive in many African languages such as Ibibio demonstrates that it must be accounted for phonologically rather than merely as phonetic implementation (Connell, 2001). Finally, downdrift has most commonly been treated as synonymous with automatic downstep, the progressive lowering of H tones when separated by L tones (Downing & Rialland, 2016). However, Connell (2001) states that downdrift is distinct from automatic downstep as found in classic terracing languages. In Mambila, for instance, the H tone is lowered by the preceding L tone, but subsequent H tones rise back to their original height, a pattern that lacks permanent terracing. This distinguishes downdrift as a local phonetic event rather than a register shift.

2.2.3 Phonetic Realisation

As mentioned in the previous sections, AM Theory emphasises the separation of phonology and phonetics, with phonetic implementation serving as the interface between abstract tonal representations and their phonetic realisation. It encompasses alignment, scaling, interpolation, and underspecification, though intonation may also influence other elements, such as duration and quality of segments (Arvaniti, 2022). The following sections discuss each aspect of phonetic realisation individually.

2.2.3.1 Alignment

Alignment refers to the timing of tonal targets relative to segmental landmarks, such as syllable boundaries or stressed syllables. Tones are realised at specific points in speech, distinct from models that emphasise pitch movements. Alignment gives rise to segmental anchoring

(Arvaniti et al., 1998), in which tonal targets, such as L and H, often align with phonetic landmarks, such as the onset or nucleus of a syllable. Furthermore, alignment patterns vary across languages and contexts, influenced by factors such as speaking rate, vowel length, and syllable structure.

The debate on alignment centres around whether tonal targets are absolutely aligned or proportionally aligned. Both are reported in languages. The argument for absolute alignment maintains that tones consistently align with fixed segmental landmarks, regardless of variations in speech rate or duration. For instance, in Greek, L and H tones are aligned with the onsets of accented syllables or post-accentual vowels, respectively (Arvaniti et al., 1998). In contrast, proportional alignment proposes that tonal alignment varies in relation to the duration of the anchoring segment. For example, in English, H* peaks align proportionally to the duration of accented syllables (Silverman & Pierrehumbert, 1990).

2.2.3.2 Scaling

The second aspect is scaling, referring to the relative pitch height of tonal targets, such as H and L tones, within a speaker's pitch range. It includes two primary perspectives: pitch span (measuring the range of frequencies used) and pitch level (Ladd, 1996). Different pitch scaling can significantly alter the pragmatic meaning of identical tonal sequences, such as L+H* in Catalan. At a lower pitch level, L+H* typically marks information-focus statements, expressing neutral and non-contrastive information. When the pitch peak is moderately raised, it can indicate corrective or contrastive focus, often used to reject a prior assumption or assert an alternative. At the highest end of the pitch range, L+H* is interpreted as a counter-expectational question expressing surprise or disbelief (Borràs-Comes et al., 2014). Scaling is frequently affected by linguistic, physiological, and paralinguistic factors.

Focusing on the linguistic aspects of tonal scaling, researchers have generally identified three primary influences: declination, tonal context, and tonal identity (Arvaniti, 2022). Declination refers to the systematic lowering of tonal targets along a declining baseline that is specific to each speaker and invariant at a given time (Pierrehumbert, 1980). This baseline is characterised by a stable minimum value and slope, though the effect of declination can be suspended in certain contexts, such as in questions (Yuen, 2007), and reset across phrasal boundaries (Truckenbrodt, 2002). Tonal context involves interactions between tones, such as upstep and downstep, which modify tonal scaling due to the influence of neighbouring tones. For instance, in English, a preceding H tone in a vocative chant can upstep the subsequent L% boundary tone (Beckman & Pierrehumbert, 1986). However, such interactions are not universal, as demonstrated in Greek, where bitonal accents and edge tone combinations do not cause upstep or downstep (Arvaniti & Baltazani, 2005). Finally, tonal identity influences how L and H tones scale under varying pitch spans; L tones tend to become lower and H tones higher when the pitch span expands (Gussenhoven & Rietveld, 2000). There are also consistent scaling differences between comparable tonal events. For example, H*+L is typically scaled higher than H* in some languages (Arvaniti & Baltazani, 2005).

2.2.3.3 Interpolation

Interpolation refers to how pitch transitions between tonal targets, H and L tones, within pitch contours. Although pitch interpolation is traditionally assumed to follow linear paths for simplicity, particularly in AM Theory, growing evidence suggests that it often exhibits nonlinear shapes, such as concave (dipping) or convex (rising) curves, depending on tonal context and linguistic factors. To account for this, researchers such as Barnes et al. (2012) have developed methods such as the Tonal Centre of Gravity (TCoG) to capture perceptual differences in nonlinear contour shapes. Furthermore, the fact that tones have duration, as seen

in plateau-like realisations of H tones, challenges the traditional notion of distinct turning points. Because the functions of plateaus compared to peaks differ across languages (appearing interchangeable in some but semantically distinct in others), the debate persists as to whether interpolation should remain a phonetic detail or be integrated into formal phonological models. While opponents argue that treating interpolation as nonlinear unnecessarily complicates phonological theory, proponents contend that overlooking these complexities simplifies the phonetic and perceptual aspects of intonation.

2.2.3.4 Underspecification

Underspecification means that not all tonal elements in a pitch contour or melody are explicitly assigned specific tonal targets or F0 frequencies during speech production. In AM Theory, melodies are often sparsely specified, with only key tonal targets (e.g., high or low points) being explicitly represented. The intermediate values between these targets are interpolated, rather than individually defined. For example, in Japanese, Pierrehumbert and Beckman (1988) demonstrated that the pitch contours of accent phrases could be modelled using just one H and one L target, with the slope of the F0 contour determined by the number of moras between them. This contrasts with full-specification models such as PENTA, which attempt to assign tonal values to every syllable, requiring the explicit representation even for interpolated values. Underspecification also applies to scenarios of tonal crowding, where there are more tonal targets than syllables available to realise them, making it difficult to implement all the tones within a limited time.

The debate around underspecification is about whether sparse (underspecified) or full specification better describes how tonal representation and realisation work. Proponents of underspecification contend that sparse representation accurately embodies the linguistic and

phonetic realities of numerous languages, including Japanese and Korean, where pitch contours can be effectively modelled without designating a specific tonal target for each syllable. They have also stressed how flexible underspecification is in dealing with tonal crowding by using methods like truncation (cutting out parts of the contour), undershoot (not fully reaching tonal targets), and temporal realignment (changing the timing of tones). As for full-specification models, proponents argue that every tonal element must be explicitly represented to align with its communicative function. This approach, however, often leads to arbitrary assumptions about tonal strength, particularly in cases of headless constituents or tonal crowding. For example, evidence from Polish shows when tonal targets compete for limited space, speakers prioritise crucial targets (like boundary tones) while truncating others (like a pitch rise).

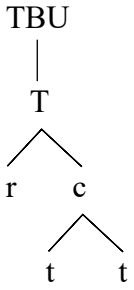
2.2.4 AM Theory in Chinese Languages: Tone Representation

Drawing on the feature geometry of the autosegmental framework, Yip (1980) addressed the distinction between pitch height (register) and pitch movement (contour). She proposed a model separating tones into two registers: upper and lower, based on the binary feature [±upper], reflecting different portions of the vocal range. Within each register, the pitch is further refined by the feature [±high], resulting in four possible tone configurations:

[+upper]	[+high]
	[-high]
[-upper]	[+high]
	[-high]

Subsequent scholars proposed further modifications to the phonological representation of tones, such as M.-C. L. Chang (1992) and Yip (1989, 2002), yet the fundamental distinction between register and contour remained unchanged. A key topic in the literature concerns the structural relationship between these two features. Two primary models have been proposed to describe

the relationship between register and contour in tonal structure. The first is the register-dominant model. Yip (1989) proposed a hierarchical relationship in which the register node dominates the contour, structurally embedding the contour feature within the register. The second is the sibling model (e.g., Bao, 1999; Duanmu, 2007), in which both register and contour are represented as parallel branches (i.e., sister nodes under a tonal root node). Bao (1999) formalised this using two nodes: the r node (register) and the c node (contour). The r node defines the relative pitch height (i.e., the register), while the c node determines the tonal movement over the duration of the tone-bearing unit. From a conceptual standpoint, the two components are distinct in function. The r node is static, indicating a stable pitch register, either H or L, within which the tone is realised. In contrast, the c node is dynamic, representing tonal movement over time. For instance, a branching c node denotes a contour, such as a falling contour that results from a sequence of high (h) followed by low (l) values, and a rising contour, which involves a shift from low (l) to high (h) values. Altogether, tones can be represented in the following structural format, leading to eight possible tonal configurations:



Four of these are level tones, maintaining a consistent pitch throughout the tone-bearing unit, such as [H, h], [H, l], [L, h], and [L, l], each indicating a steady pitch height in either the high or low register, in which the contour is non-branching. The other four patterns are contour tones, which consist of a dynamic movement in pitch. These include two falling tones [H, hl] and [L, hl], and two rising tones [H, lh] and [L, lh].

2.3 Summary

This chapter introduced different intonational models, and, in particular, examined AM Theory. It outlined the origins and core principles of AM Theory and explored how phonetic realisation mediates between phonology and speech by mapping abstract tonal structures onto acoustic features, with particular attention to alignment, scaling, interpolation, and underspecification. By applying AM Theory to tone in Chinese languages, this chapter sets the stage for the next chapter, which examines the production of question intonation in GuanM and explores how tonal and intonational patterns interact in speech.

Chapter 3 Production Study 1: Intonational Yes-No Questions in Guanzhong Mandarin in Isolation

The previous chapter introduced AM Theory, the theoretical foundation of this thesis. This chapter begins by exploring intonation in GuanM, focusing on the production of syntactically unmarked yes-no questions in isolation. The chapter is structured as follows: it first reviews previous studies on intonation concerning yes-no questions across various languages, with a specific emphasis on Chinese languages. It subsequently presents the research questions and describes the methods used in this production study, including participant recruitment, stimuli, recording procedure, data sorting, and statistical analysis. Finally, the chapter ends with the presentation of results, followed by a discussion and final thoughts.

3.1 Question Marking Cross-Linguistically

Two common strategies are used cross-linguistically to form yes-no questions. The first approach employs syntactic-lexical devices, which are explicit grammatical markers that signal a question, including question particles such as *ma* in SC and *ka* in Japanese, interrogative morphemes, such as *mi* in Turkish, word order inversion as seen in the verb-first structure of German and French, and the use of auxiliary words or phrases, such as *est-ce que* in French. The second method, commonly referred to as declarative questions, depends exclusively on intonation, in which questions have the same lexical elements and word order as declarative sentences but utilise pitch variations to indicate their questioning function. Cross-linguistically, question intonation is frequently characterised by a rising terminal pitch, with Bolinger (1951) observing that most yes-no questions had a rise somewhere in the utterance. By examining yes-no question marking across 78 African languages, Rialland (2009) argued that the rise was not always present and identified two categories of yes-no question markers. The first category

consisted of five high-pitched strategies: the cancellation or reduction of downdrift or register expansion, the raising of final H tone(s), the cancellation or reduction of final lowering, a final H tone or rising intonation, and a final HL melody. The second category included six non-high-pitched strategies: a final L tone or falling intonation, a final polar tone or mid tone, lengthening, breathy termination, the cancellation of penultimate lengthening, and open vowels.

The specific intonational patterns for unmarked yes-no questions vary across languages. Haan's (2002) study showed that in Dutch, 100% of declarative questions had a final rise (H%), which was realised at the highest point in the overall pitch range compared to other question types, and they were generally of higher pitch than statements. Research by Bibyk (2016) showed that information-seeking declarative questions in American English generally tended towards rising over falling intonation. In Manchego Peninsular Spanish (Henriksen, 2010), declarative questions exhibit two contour types: early rise, which appears more frequently in casual dialogue, signalling their interrogative status through both higher nuclear scaling and raised final pitch; and late rise, primarily found in formal read speech, depending only on higher final tonal values to mark the question-statement distinction. Specifically, the early rise contour (L* + H H* ¡H%), characterised by a nuclear H* pitch accent followed by an upstepped ¡H% boundary tone, begins its ascent near the onset of the nuclear stressed syllable and maintains a relatively shallow F0 decline from the prenuclear peak. The late rise contour (L* + H L* H%), featuring a nuclear L* accent followed by an H% boundary tone, has its rise beginning towards the end of the nuclear syllable after a substantial F0 fall from the prenuclear peak. E. Lin et al. (2013)'s cross-language study of declarative questions in tonal and non-tonal languages also demonstrated that regardless of whether a language used tone to distinguish word meanings (such as Mandarin and Taiwanese Hokkien) or not (such as English and Arabic), declarative yes-no questions consistently showed higher pitch and intensity than statements. While non-

tonal languages showed more distinct pitch rises for questions, even tonal languages maintained this distinction despite the competing demands of lexical tone.

However, the terminal rise is not the only intonational strategy for marking yes-no questions. Bengali yes-no questions exhibit a rise-fall contour, in which the nuclear-stressed syllable has a low pitch, followed by a gradual rise to H near the end of the phrase, and then a rapid fall at the very end (Hayes & Lahiri, 1991). Similarly, in Puerto Rican Spanish, the rise-fall contour appears in three distinct realisations for yes-no questions, which are ¡H*L%, the most common pattern, consisting of an extra-high pitch accent followed by a fall; L*HL%, which begins with a low pitch accent followed by a rise-fall at the boundary, typically signalling disbelief or counter-expectation; and H+L*L%, which starts high and falls, often conveying positive epistemic bias (Armstrong, 2017). Savino (2012) conducted a typological study on 15 varieties of Italian yes-no questions and found that 7 varieties (Turin, Venice, Parma, Bari, Naples, Catanzaro, and Palermo) exhibited rise-fall contours, while 3 additional varieties (Genoa, Florence, and Rome) exhibited mixed patterns where the accentual rise could be followed by either a falling or rising boundary. As a consequence, Savino concluded that a terminal rise was not the most typical intonational strategy in Italian questions; instead, a rise on the nuclear syllable (accentual rise), generally followed by a fall, was a widespread intonational feature for signalling questions across Italian varieties, not just in Southern accents.

Intonation may not be a necessary marker for declarative questions. Chłopicki's (2019) study of Polish student conversations discovered that declarative questions in Polish had rising intonation as a common but not obligatory phonetic marker of interrogativity, with no formal question markers used. Hennoste et al.'s (2019) research on Estonian declarative questions also found that intonation alone was not a reliable indicator of a questioning function. Intonational

patterns were highly variable, with some questions displaying a slight final rise while others showed steeper declination than statements. The authors argued that speakers relied more heavily on pragmatic factors such as epistemic status and conversational context.

Before turning to a detailed examination of question marking in SC and Chinese dialects, a summary table of the yes-no question marking strategies discussed earlier is presented below.

Table 3.1 Yes-no question marking cross-linguistically

Intonational Pattern	Languages	Notable Features
Terminal Rise / Final High Tone	Dutch	• 100% of declarative questions have final rise (H%) • Realised at the highest point in overall pitch range • Generally higher pitch than statements
	American English	• Information-seeking declarative questions prefer rising intonation over falling
	Manchego Peninsular Spanish (Late rise)	• L* + H L* H% pattern • Found mainly in formal, read speech • Rise begins towards end of nuclear syllable
Early Rise	Manchego Peninsular Spanish	• L* + H H* ¡H% pattern • More frequent in casual dialogue • Ascent begins near onset of nuclear stressed syllable • Shallow F0 decline
Rise-Fall Contour	Bengali	• Low pitch on nuclear-stressed syllable • Gradual rise to H near phrase end • Rapid fall at the end
	Puerto Rican Spanish	• ¡H*L%: Extra-high pitch followed by fall (most common) • L*HL%: Rise-fall boundary (signals disbelief) • H+L*L%: High start, falls (positive epistemic bias)

	Italian varieties (Turin, Venice, etc.)	• Accentual rise on nuclear syllable followed by fall • Widespread across Italian dialects
Mixed / Variable Patterns	Italian varieties (Genoa, Florence, Rome)	• Accentual rise followed by either falling or rising boundary
	Polish	• Rising intonation common but not obligatory • No formal question markers
	Estonian	• Intonation not a reliable indicator • Some questions rise slightly; others remain flat or fall
Higher Overall Pitch	Tonal languages (Mandarin, Taiwanese Hokkien)	• Despite tones, maintain a higher pitch for questions
	Non-tonal languages (English, Arabic)	• Tend to use more distinct pitch rises for questions
	African languages (Rialland, 2009)	• Cancellation/reduction of downdrift or register expansion • Cancellation or reduction of final lowering • Raising of final H tone(s) • Final H tone or rising intonation • Final HL melody
Non-High-Pitched Strategies	African languages (Rialland, 2009)	• Final L tone or falling intonation • Final polar or mid tone • Lengthening • Breathy termination • Open vowels • Cancellation of penultimate lengthening

The prosodic and intonational patterns of yes-no questions in SC and Chinese dialects have been extensively studied. Research has focused on debates surrounding local and global trends in question prosody, the interaction between lexical tones and intonation, and the definition of Chinese boundary tones. Chao (1968) was one of the pioneers in Chinese tone and intonation studies. He extensively discussed tone, intonation, and their interaction in Chinese languages and proposed a metaphor comparing the relation between syllabic tone and sentence intonation to small ripples riding on larger waves (though, at times, the ripples became more prominent

than the waves themselves), where the resulting intonation was an algebraic sum of both elements. Specifically, when positive values aligned with other positive values, the result was an enhanced positive value; when a positive value encountered a negative value, the mathematical addition resulted in subtraction. This meant that when two rising movements coincided, the outcome was an even stronger rise, whereas when a rising and a falling movement interacted, their effects partially cancelled out, much like arithmetic subtraction. Chao (1933) also identified two types of tone and intonation addition patterns based on observations. The first type, simultaneous addition, referred to tones that were the algebraic sum of the original lexical tone and the sentence intonation, or in other words, the resultant of both. The second type, successive addition, described a rising or falling intonation that was applied sequentially after the completion of lexical tones. Chao formalised how rising and falling intonation modified lexical tones using the following tonal transformations in SC:

Table 3.2 The modification of tones by intonation proposed by Chao (1933)

Original Tonal Value	Change with Rising Intonation	Change with Falling Intonation
T1 (55)	56	551
T2 (35)	36	351
T3 (214)	216	2141
T4 (51)	513	5121

In the table above, 6 denotes extra-high pitch. In rising intonation, T4 acquires an upward pitch prolongation, while in the case of falling intonation, a downward pitch prolongation is grafted onto all four tones. Consequently, both tone shape and tonal value are altered to such an extent that these tones become less recognisable (Shen, 1989).

The determination of whether Chinese yes-no questions are marked by rising or falling intonation remains an active area of research. Gårding (1987) applied a grid structure to model intonation in Chinese. In general, yes-no questions showed a rising intonation contour, particularly over the predicate part of the sentence, while statements mostly followed a falling trend. However, when X. Xu et al. (2018) attempted to model morphologically unmarked yes-no questions in SC, they did not observe a rising intonation pattern that accelerated towards the end, as was previously assumed. Instead, yes-no questions had a nearly horizontal or slightly declining slope, which was significantly flatter than the steeper downward pattern observed in declarative statements. A clear three-pattern structure was apparent: a rising trend at the beginning of the sentence, a moderate downward slope in the middle, and a return to a more horizontal pattern at the end.

Although scholars generally agreed that yes-no questions in SC exhibited rising intonation and declarative sentences exhibited falling intonation (similar to some other languages), debate persisted over the temporal characteristics of the rising contour. Specifically, researchers were unsure whether yes-no questions in Mandarin were primarily marked by global pitch raising throughout the utterance or by local modifications. Shen (1989) found that in yes-no questions, F0 was raised not only on the final syllable but throughout the entire utterance. Similarly, Ni and Kawai (2004) noted that the overall pitch of unmarked yes-no questions tended to be higher than that of a statement. However, M. Lin (2004) argued that while global pitch raising might occur, the register and slope of the F0 curve at the final syllable were more crucial for distinguishing questions from statements. Huang and Zhang (2019) examined questions with and without the final question particle [ma] ‘吗’ and found that unmarked yes-no questions compensated with greater pitch raising, lengthened duration, and higher intensity on the final syllable.

Furthermore, P. Jiang and Chen (2011) indicated that both global pitch level rising (across the entire sentence) and localised pitch raising at both edges of the intonational phrase contributed to signalling questions. F. Liu and Xu (2005) also found that yes-no questions in Mandarin followed an exponential or double-exponential pitch raising pattern extending throughout the sentence. Their later study (F. Liu & Xu, 2007), which focused on Mandarin, further illustrated that question intonation involved a global pitch increase beginning from the focused word, rather than just a final rising contour. Yuan (2006) identified three mechanisms of question intonation in SC: raising the entire phrase curve, strengthening sentence-final tones, and implementing a tone-dependent mechanism that reduces the steepness of falling final tones while increasing the steepness of rising final tones.

The interaction between lexical tones and intonation in SC shows that intonation modifies but preserves lexical tones, creating a layered relationship where question intonation affects how tones are realised acoustically without fundamentally altering tonal categories; this relationship is best understood as intonation being superimposed upon the tones (Ho, 1977). Shen (1989) observed that while intonation modified the F0 trajectory of lexical tones, their basic contour shapes remained intact: rising tones in the sentence-final position exhibited larger pitch movement in questions; the endpoints of T2 and T3 rose much higher in questions than in statements; and T4 remained relatively stable. In addition, M. Lin (2004) discovered that for T1, T2 and T4, the register of both starting and ending points was higher in questions than in statements. For T3, the rising portion of the F0 curve in questions was more prominent, distinguishing it from its typically low realisation in statements. Ni and Kawai (2004) also reported tone-dependent effects at the sentence-final position. The F0 peak of the rising portion of T1 and T3 was raised significantly, expanding the tonal range. In contrast, both F0 peaks

and valleys were raised for T1 and T4, narrowing the tonal range while maintaining the raised contour.

Regarding the intonation of yes-no questions in different dialects in China, C. Zhang (2018) used monosyllables as intonational phrases and claimed that in Tianjin Mandarin, a variety of Jilu Mandarin, yes-no questions had a higher pitch register across the intonational phrase and a high floating boundary tone, which did not alter the lexical tone contour but instead intensified the ascent of rising tones and reduced the descent of falling tones. H. Jiang (2019) confirmed Chang's (1958) observation in the Chengdu dialect (a variety of Southwestern Mandarin) that yes-no questions could be classified as having either a rising or falling tune based on the final syllable's tone after undergoing perturbation. He also found that yes-no questions generally had a higher pitch and steeper contour slope on the final syllable, with the lexical tone remaining identifiable while influencing the degree of pitch variation. Questions in the Changsha dialect, a dialect of New Xiang Chinese, tend to have an overall higher pitch than statements, although the initial pitch level of questions does not always begin at a higher point than that of statements (Shen, 1991).

Unlike most languages that mark the distinction between statements and questions at the sentence-final boundary or the whole-utterance level, Bai, a Sino-Tibetan language spoken in the southwestern region of China, primarily in Yunnan Province, establishes the difference between questions and statements in the sentence-medial position. Bai speakers differentiate yes-no questions from statements through prosodic alterations of sentence-medial constituents, including increased duration, expanded pitch ranges, and raised pitch maxima and minima, regardless of focus conditions. Yet these modifications are tone-dependent such that the pitch

span of sentence-medial constituents for rising and falling tones is expanded and the pitch minimum is raised for level and rising tones (Z. Liu et al., 2018).

Yes-no questions in Cantonese provide a different perspective on the interaction between tone and intonation. In this case, intonation alters the final lexical tone contour shape at the last syllable, unlike in most Mandarin dialects. J. K.-Y. Ma et al. (2006a) found that in the final position of questions, all six Cantonese tones had a rising F0 contour, with tones 25, 21, 23, and 22 overlapping in their F0 contours, resembling tone 25, regardless of their original tonal shape. In addition, questions exhibited an overall higher F0 than statements, with the greatest difference in F0 between the two intonation types occurring at the phrase-final position.

3.2 Research Questions

The difference between intonation and tune must first be clarified. Both refer to sentence prosody. However, the term intonation refers to sentence-level prosody in a broader sense, encompassing F0 properties as well as other related sentence-level prosodic components, such as duration. In contrast, a tune is an abstract representation of sentence-level prosody, independent of specific linguistic content (Hayes & Lahiri, 1991), conveying the intonational information. For example, the tune of expressing surprise in English is LHL, as shown in (8), regardless of different tone associations with segments, with its association rules illustrated in (9).

(8) Surprise tune in English

(9) Association rule for the surprise tune (Hayes & Lahiri, 1991)

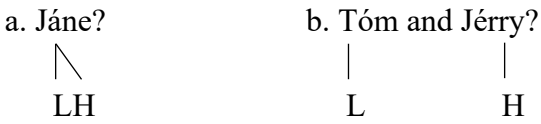
a. Associate H to the main stressed syllable.

b. Associate the first L to syllables preceding the main stress, if there are any.

c. Associate the second L to all syllables following the main stress; if there are none, associate L to the same syllable as the H tone.

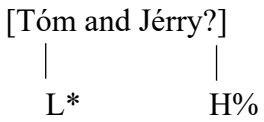
The tune for yes-no questions in English is LH, as shown in (10), regardless of the varying utterance length.

(10) Question tune in English



The boundary tone of the question tune in English shown in (10)b can be hypothetically represented as follows.

(11)



Yes-no questions in GuanM, much like in SC, can be categorised into syntactically marked and intonational questions. Syntactically marked questions are often accompanied by sentence-final particles, such as [ma] ‘吗’, [me] ‘么’, and [pa] ‘吧’， explicitly marking interrogativity. Besides sentence-final particles, GuanM is distinguished from SC by [tei31 si55] ‘得是’, an interrogative adverb that indicates the speaker’s degree of certainty regarding the proposition being questioned. It functions as a question focus marker that conveys a positive epistemic bias. When used in yes-no questions, it shows that the speaker has already formed a tentative judgement that the proposition is likely to be true, yet still anticipates that the hearer might hold

a different view. The speaker, therefore, seeks confirmation or affirmation rather than new information. For instance, in 他得是你学校的老师？[tʰa31 tei31 si55 ni52 xio24 xiau55 ti lau52 ʃi31] 'He must be a teacher at your school, right?', the use of [tei31 si55] expresses a confirmatory tone, signalling the speaker's expectation that the proposition is true (L. Zhao, 2019).

Little formal research has examined intonational yes-no questions in GuanM. Therefore, the primary aim of this chapter is to investigate the intonation patterns of syntactically unmarked yes-no questions in GuanM. Using disyllabic sequences with identical tones (T1T1, T2T2, and T4T4) as intonational phrases, this chapter particularly focuses on the interaction between lexical tones and intonation, specifically how the T1T1 tone dissimilation rules function within the interrogative tune. As discussed in the previous section, GuanM has two tonal dissimilation rules: T1T1 (HL_{Lr}→LH/_ HL_{Lr}) and T3T3 (HL_{Hr}→HL_{Lr} /_ HL_{Hr}). The latter is excluded from the present analysis because it appears to be undergoing diachronic change. The status of the T3T3 alternation remains unsettled: older speakers tend to preserve the dissimilatory pattern, whereas younger speakers are more likely to retain the original tone values. Moreover, unlike the T1T1 rule, in which the contour shape of the first tone changes, the T3T3 rule primarily involves register lowering rather than contour modification. For this reason, T3T3 is treated here as background and is not included in the present analysis.

3.3 Methods

3.3.1 Participants

Eight native speakers of GuanM participated in this study (three females, five males; age range: 20 to 45 years old; mean age: 31.88 years old). They were from Xi'an (Shaanxi), China, and

spoke GuanM daily. None of the participants reported any hearing or speaking disorders, and they gave their written informed consent prior to the experiment.

3.3.2 Stimuli

In total, 45 disyllabic words were selected: fifteen from each of the T1T1, T2T2, and T4T4 sequences. All selected words consisted entirely of sonorants to avoid the F0 perturbations typically caused by obstruents and to ensure more accurate alignment and annotation. Table 3.3 lists all the stimuli that were used in this study.

Table 3.3 Stimuli used in Production Study 1

T1T1	**T2T2**	**T4T4**
乌鸦	麻油	浪漫
u31ia31	ma24iou24	laŋ55man55
‘crow’	‘sesame oil’	‘romance’
鸳鸯	绵羊	面料
yan31iaŋ31	mian24iaŋ24	mian55liau55
‘mandarin duck’	‘sheep’	‘fabric’
妖孽	棉麻	面貌
iau31ŋiɛ31	mian24ma24	mian55mau55
‘evildoer’	‘linen fabric’	‘face’
落叶	蓝莓	样貌
luo31iɛ31	lan24mei24	iaŋ55mau55
‘fallen leaves’	‘blueberry’	‘appearance’
音乐	羊毛	奥运
ien31io31	iaŋ24mau24	ŋau55yen55
‘music’	‘wool’	‘Olympic games’
录音	农民	外卖
lu31ien31	nuŋ24mien24	uæ55mæ55
‘recording’	‘farmer’	‘takeout’
弯月	南门	命运
uan31yɛ31	nan24men24	miŋ55yen55
‘crescent moon’	‘south gate’	‘destiny’

阅历	杨梅	外贸
yɛ31li31	iaŋ24mei24	uæ55mau55
‘experience’	‘bayberry’	‘foreign trade’
月末	良民	命令
yɛ31mo31	liaŋ24mien24	miŋ55liŋ55
‘end of month’	‘good citizen’	‘order’
目录	谣言	利用
mu31lu31	iau24ŋian24	li55yŋ55
‘catalogue’	‘rumour’	‘use’
优越	毛驴	另类
iou31yɛ31	mau24ly24	liŋ55luei55
‘superior’	‘donkey’	‘alternative’
沐浴	煤油	暧昧
mu31y31	mei24iou24	ŋæ55 mei55
‘bathing’	‘kerosene’	‘ambiguous’
烟叶	邮轮	命案
ian31iɛ31	iou24lyen24	miŋ55ŋan55
‘tobacco’	‘cruise’	‘homicide’
英烈	牦牛	议论
iŋ31liɛ31	mau24ŋiou24	i55lyen55
‘heroes’	‘yak’	‘discussion’
邀约	鱿鱼	义卖
iau31ŋio31	iou24y24	i55mæ55
‘invite’	‘squid’	‘charity sale’

3.3.3 Procedure

Before the study, ethical approval was obtained from the university’s research ethics committee (CUREC 1A; reference number R74409/RE001). All audio recordings were made in a quiet room within the local community centre using a Samson Meteor Mic connected to a personal Lenovo laptop; the recordings were processed via Audacity software at a sampling rate of 44.1 kHz.

The stimuli were presented to participants using Microsoft PowerPoint, with each slide displaying only one target disyllabic word, followed by either a full stop “。” or a question

mark "？". Each target disyllabic word was repeated three times in a pseudorandomised order to ensure that participants' responses were not influenced by encountering the same stimuli too often. Participants were instructed to imagine a conversational scenario for each slide. For words followed by a full stop "。", participants were asked to produce the word with declarative intonation, as if confirming or identifying it in response to someone pointing to or mentioning it, similar to responding "Yes, that's [word]." For words followed by a question mark "？", they were asked to produce the word with interrogative intonation, as if they had heard someone say it unclearly and were seeking verification, similar to asking "Did you say [word]?". For example, when presented with "麻油 [ma24 iou24]。" on the screen, participants would be expected to produce it with declarative intonation, as if confirming "That's sesame oil". When presented with "麻油 [ma24 iou24]？", they would be expected to say it with interrogative intonation, as if asking "Sesame oil? Is that what you meant?".

Before the main experiment, participants were given time to familiarise themselves with the task and completed 10 practice trials using words that were not included in the experimental stimuli. After completing the main task, participants also recorded all stimulus words twice to establish a baseline. To maintain experimental consistency, these baseline tokens were presented in the same format as the main task: each word appeared individually on a slide, and the presentation order was pseudorandomised to mitigate order effects. Unlike the main task, these tokens were produced in isolation without conversational context or punctuation cues, serving as disyllabic reference forms[8]. It should be noted that while these baseline trajectories

[8]Although monosyllabic citation forms are traditionally used as a baseline in tonal studies, this study uses disyllabic reference forms instead. Since the intonational tunes under investigation are realised over a disyllabic domain, the reference data must match this structure. Monosyllables would not be suitable, as isolated tones tend to differ from tones in connected speech in duration and scaling. These disyllabic reference forms were included

provide a necessary qualitative reference for visualising F0 shifts, they were not included in the formal statistical models, which were restricted to the comparison between declarative and interrogative intonations to directly quantify the phonetic interaction between sentence type and tone.

3.3.4 Annotation and Alignment

For each speaker, two of the three repetitions were selected for analysis. The selection was made primarily at random; however, repetitions with obvious recording problems, annotation difficulties, or unclear F0 contours were excluded before selection, which ensured that the analysis was based on consistent and clearly analysable audios while minimising the influence of outliers or potential production errors. This selection resulted in a total of 180 recordings per speaker (3 tones × 15 disyllables × 2 tunes × 2 repetitions). Each individual spectrogram was visually inspected and subsequently annotated using Praat (Boersma & Weenink, 2023). To maintain consistency in the annotations, the onset of each utterance was marked at the beginning of the F0 and the appearance of the F1, F2, and F3 formants, while the offset was marked by the disappearance of the F0, F1, F2, and F3 formants. The audio files were then annotated with two tiers: the syllable tier and the entire phrase tier.

3.3.5 Measurement

The acoustic measurements from the audio trials included duration, mean intensity, and register change at both the phrase level and the individual syllable level. Duration was measured as the length of the entire word phrase as well as each individual syllable. Register change was

for descriptive purposes only, providing a visual reference for the F0 trajectories without intonational overlay. They were not included in the formal statistical analyses, which compared declarative and interrogative conditions directly. Since the research question concerns how question intonation interacts with tonal realisation, the statement condition serves as the baseline against which question effects are measured (e.g., Yuan (2011) conducted his study where only two intonations were directly compared).

examined through several parameters, including mean F0, maximum F0, minimum F0, and F0 range, as well as the alignment of the maximum and minimum F0 for both the entire phrase and each syllable.

3.3.6 Statistical Analyses

The normalised F0 contours in both intonations were calculated by extracting ten evenly distributed F0 points from each syllable for each speaker to analyse the F0 trajectories across these tones over time. The F0 values were then normalised using semitone conversion (J. W. Zhang, 2018)[9], calculated relative to each speaker's average F0 across all their utterances. The formula for this conversion was as follows:

$$ST_{F0} = 12 \times log_2(\frac{F0}{mean\ F0})$$

Following normalisation, the rate of F0 change was calculated to quantify the overall direction and speed of pitch movement over time. It is important to note that the purpose of this analysis was not to model the precise shape of the tonal contour within each syllable, which is often nonlinear in Chinese languages, but rather to extract the overall slope as a measure of how rapidly and in which direction F0 changes across the measured duration, which allows us to examine how the overall pitch trajectory shifts under different intonational conditions; that is, whether question intonation raises, lowers, or alters the angle of the pitch movement compared to statements. Such a method is motivated by the phonetic implementation of AM Theory including interpolation (Arvaniti, 2022), emphasising the pitch trajectory between tone targets, as mentioned in Section 2.2.3.3.

[9] J. W. Zhang (2018) systematically compared 16 tone normalisation methods and concluded that ST-AvgF0 (semitone transformation relative to each speaker's mean F0) is the most effective method for tonal variation studies, as it preserves phonemic contrasts, minimises anatomical differences, and retains sociolinguistic variation across speakers.

For this purpose, linear regressions were computed separately for each syllable (10 evenly distributed points) and for the entire phrase (20 evenly distributed points) using the lm function in R (R Core Team, 2024). The linear regression equation used was:

$$y = kx + b$$

where k (the slope) represents the rate of pitch change over time, and b (the y-intercept) indicates the starting pitch value at the beginning of the syllable.

Furthermore, generalised additive mixed models (GAMMs) were fitted in R (R Core Team, 2024) using the mgcv package (Wood, 2011) to capture nonlinear F0 and variability across speakers and items. Semitone-normalised F0 was modelled using smooth functions of normalised time, with an interaction between Tone and Tune and random effects of Speakers and Items.

For the statistical analysis, linear mixed-effects models were fitted using the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2024) to examine the interaction between Tune and Tone across various acoustic variables, including Duration, Mean Intensity, Mean F0, Maximum F0, Minimum F0, F0 range, Maximum F0 alignment, and Minimum F0 alignment at both whole-phrase and syllable levels, including an interaction term of Syllable across three tonal categories: T1T1, T2T2, and T4T4. Random effects included Speaker, Repetition, and Item. Post-hoc pairwise comparisons were further conducted using the emmeans package (Lenth, 2024).

3.4 Results

3.4.1 Normalised Representation of F0 Curves

Figure 3.1 to Figure 3.3 present examples of waveforms for each of three tones in a statement (on the left) and in a question (on the right) produced by the same speaker. These figures also include spectrograms displaying the F0 contours, together with the annotated files that show the detailed alignments and annotations.

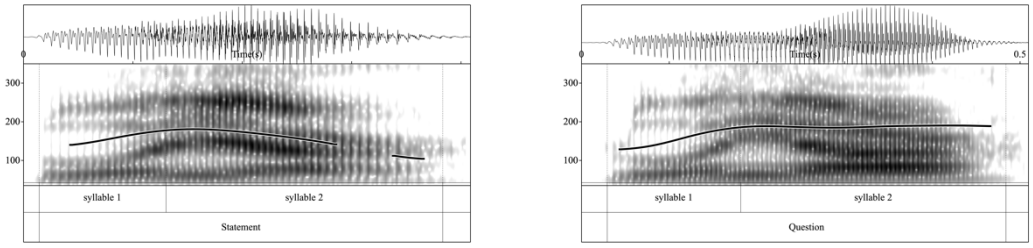


Figure 3.1 Statement tune (left panel) and question tune (right panel) of T1T1 produced by the same speaker

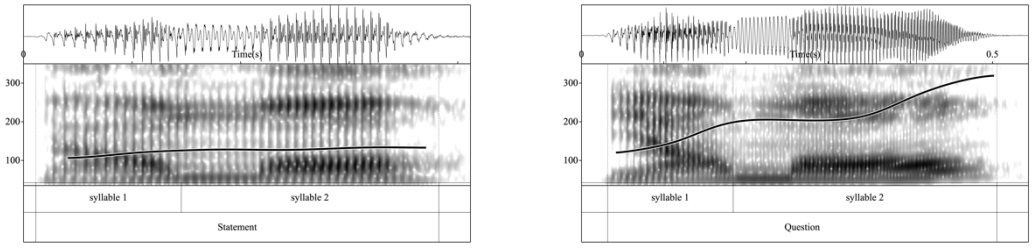


Figure 3.2 Statement tune (left panel) and question tune (right panel) of T2T2 produced by the same speaker

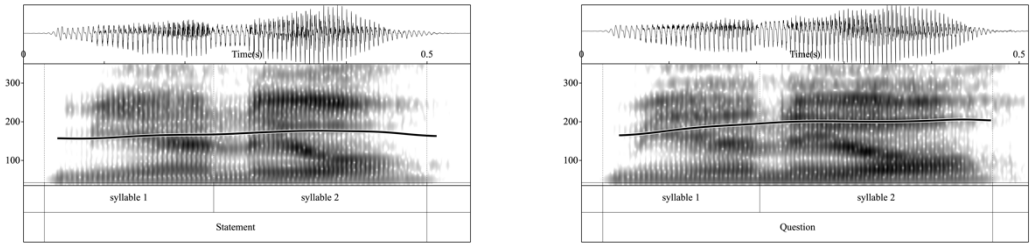


Figure 3.3 Statement tune (left panel) and question tune (right panel) of T4T4 produced by the same speaker

Figure 3.4 below presents the time-normalised representation of mean F0. A consistent pattern was found across all three tones: the F0 curves of the question tune were consistently positioned above those of the statement tune, indicating a higher overall F0 for the question tune, while the curves of the disyllabic reference form tended to fall in between the curves for the question tune and the statement tune. Furthermore, in the case of the T1T1 question tune, the first syllable (the first 10 points along the x-axis) maintained its dissimilation contour, shifting from a falling tone to a rising tone, while the second syllable (the last 10 points) did not exhibit as significant a drop as observed in its declarative counterpart. For T2T2, the F0 curve of the second syllable in the question tune showed a clear rise compared to its declarative form, suggesting the potential function of a high boundary tone. As for T4T4, the F0 curve of the question tune rose even further, thus emphasising a higher pitch associated with the interrogative tune. Another important observation was in T2T2 and T4T4 the question and reference-form contours intersected in which the question tune began below the reference form before crossing above it. No such intersection occurred in T1T1, where question tune remained above the reference form throughout, diverging without crossing.

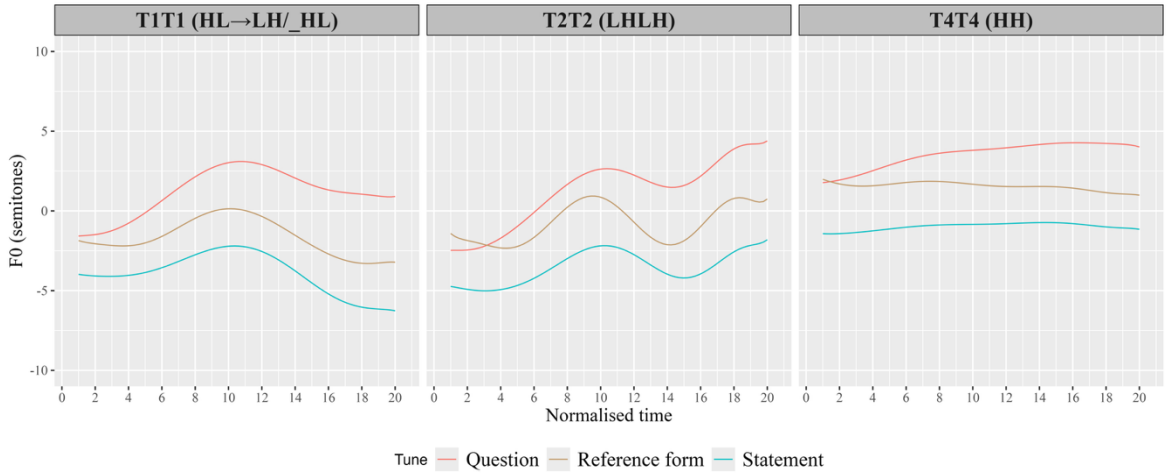


Figure 3.4 Time-normalised mean F0 contours of disyllabic words across three tones

3.4.1.1 F0 Change Rate

Regarding the F0 change rate, which was calculated according to the slopes and intercepts, both intercepts and slopes for the question tune were larger than they were for the statements at the level of the entire phrase (Table 3.4). For T1T1, at the whole-phrase level, the question tune had a positive slope of 0.13, indicating a slight rising tendency, whereas the statement had a negative slope of -0.12, indicating a falling tendency. The intercept for questions was consistently higher than it was for statements across all the syllables and tones (Table 3.5), indicating that the pitch (or F0 starting point) in questions was generally higher than it was in statements. The slopes were consistently positive for both tunes for the first syllable across all three tones, including the dissimilation contour of T1T1, thus suggesting an overall trend towards a rising pitch. The slopes were also larger in questions, indicating a faster rate of pitch change in questions compared to statements. In the second syllable, the slope was negative for T1T1 in both tunes, reflecting the falling nature of T1. However, the slope for the statement (-0.50) was steeper (i.e., more negative) than was that for the question (-0.27), implying that the pitch in T1T1 statements fell more rapidly than it did in questions. For T2T2 and T4T4, the

slopes were again larger in questions, further supporting the pattern of a faster pitch change in questions compared to statements.

Table 3.4 Intercept and slope with SD in parentheses for three tones

Tone	Tune	Intercept (SD)	Slope (SD)
T1T1	Q	-0.30 (2.68)	0.13 (0.20)
	S	-2.73 (2.83)	-0.12 (0.28)
T2T2	Q	-2.33 (2.25)	0.34 (0.15)
	S	-4.80 (2.22)	0.12 (0.14)
T4T4	Q	2.20 (1.91)	0.13 (0.13)
	S	-1.23 (2.24)	0.02 (0.10)

Table 3.5 Intercept and slope with SD in parentheses by tone, syllable and tune

Tone	Sy.	Tune	Intercept (SD)	Slope (SD)
T1T1	Sy.1	Q	-2.69 (3.09)	0.57 (0.43)
		S	-4.66 (3.17)	0.23 (0.43)
	Sy.2	Q	5.90 (5.61)	-0.27 (0.32)
		S	3.20 (9.48)	-0.50 (0.67)
T2T2	Sy.1	Q	-3.80 (2.83)	0.64 (0.36)
		S	-5.79 (2.76)	0.33 (0.32)
	Sy.2	Q	-1.46 (4.22)	0.27 (0.28)
		S	-4.37 (3.16)	0.09 (0.19)
T4T4	Sy.1	Q	1.54 (1.94)	0.25 (0.21)
		S	-1.54 (2.37)	0.08 (0.21)
	Sy.2	Q	3.72 (2.68)	0.03 (0.17)
		S	-0.28 (3.31)	-0.04 (0.18)

To further verify the differences between the slopes and intercepts of the questions and the statements, linear mixed effect models were applied to both whole-phrase level and at the syllable level for the slope and intercept analyses. These models included fixed effects for Tone, Tune, and Syllable in the syllable-level comparisons, while Speaker was treated as a random effect. Post-hoc tests were also performed, using the phia package (De Rosario-Martinez, 2012) in R (R Core Team, 2024). The results revealed that the main effects of Tone, Tune, and

Syllable (at the syllable level) were significant in both models, and interactions between Tone and Tune, as well as between Tone and Syllable (at the syllable level), were also significant ($p < .05$). The post-hoc tests showed that the intercepts and slopes differed significantly between questions and statements across all tones at both whole-phrase level and at the syllable level (all ps $< .05$). The complete post-hoc results are presented in Table 3.6 at the whole-phrase level and Table 3.7 at the syllable level.

Table 3.6 Post-hoc pairwise comparison result for intercept and slope at the whole-phrase level (Production Study 1)

	Intercept					**Slope**				
	β	SE	df	χ^2	*p*	*β*	SE	df	χ^2	*p*
T1T1	2.43	0.20	1	143.25	< .001	0.25	0.01	1	289.23	< .001
T2T2	2.47	0.20	1	148.21	< .001	0.22	0.01	1	211.10	< .001
T4T4	3.43	0.20	1	285.05	< .001	0.10	0.01	1	47.61	< .001

Table 3.7 Post-hoc pairwise comparison result for intercept and slope at the syllable level (Production Study 1)

		Intercept					**Slope**				
		β	SE	df	χ^2	*p*	*β*	SE	df	χ^2	*p*
T1T1	Sy.1	1.98	0.38	1	27.45	< .001	0.35	0.03	1	133.65	< .001
	Sy.2	2.70	0.38	1	51.1	< .001	0.23	0.03	1	60.23	< .001
T2T2	Sy.1	1.98	0.38	1	27.61	< .001	0.32	0.03	1	110.64	< .001
	Sy.2	2.91	0.38	1	59.60	< .001	0.19	0.03	1	37.98	< .001
T4T4	Sy.1	3.07	0.38	1	66.19	< .001	0.17	0.03	1	33.07	< .001
	Sy.2	4.00	0.38	1	112.07	< .001	0.06	0.03	1	4.66	.0309

3.4.1.2 GAMMs

This section provides the results obtained from the GAMMs. Figure 3.5 illustrates the differences in pitch contours among statement, question, and reference forms across the three

tones. In all three tones, the pitch contour for reference forms was usually positioned above the F0 tracks of both questions and statements, with minimal overlap in their confidence intervals (CIs). For T1T1, questions generally exhibited a higher pitch, characterised by a gradual decline in the second syllable. Conversely, statements peaked initially and then declined dramatically throughout in the second syllable. Questions displayed a rising trend over time, while statements showed a declining trajectory. The pitch contour for questions rose steadily across both syllables in T2T2 and maintained a higher average pitch compared to statements, which showed a more moderate rise in the second syllable. Questions and statements mainly differed in their pitch levels for T4T4, with questions consistently having a higher pitch across both syllables. Questions displayed a rising trend in the second syllable, whereas statements remained relatively flat. The non-overlapping CIs between statements and questions across all three tones indicated significant differences in both pitch levels and the shapes of the pitch contours in the two types of utterances.

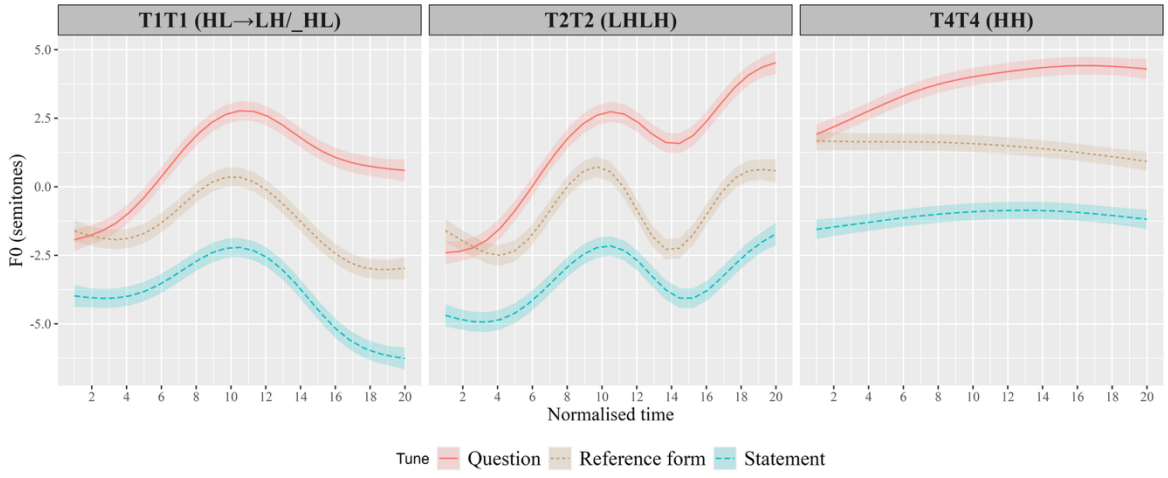


Figure 3.5 Time-normalised smoothed F0 trajectories of disyllabic words for T1T1, T2T2, and T4T4, predicted by GAMMs with 95% confidence intervals

3.4.2 Duration

At the whole-phrase level (on the left of Figure 3.6), significant main effects were observed for Tune, Tone, and their interaction (Tune: $\chi^2(3) = 291.43$, $p < .001$; Tone: $\chi^2(4) = 151.19$, $p < .001$; Tune x Tone: $\chi^2(2) = 137.36$, $p < .001$). Post-hoc analyses indicated that question tunes generally had longer durations compared to statement tunes. Specifically, the question tune duration was significantly longer for T1T1 ($\beta = 80.52$, $t(1390) = 17.32$, $p < .001$) and T4T4 ($\beta = 21.96$, $t(1390) = 4.72$, $p < .001$), although the difference for T2T2 was not statistically significant ($\beta = 5.42$, $t(1390) = 1.17$, $p = .244$).

At the syllable level (on the right of Figure 3.6), significant main effects and interactions were observed for Tune, Tone, and Syllable on Duration. The main effect of Tune was highly significant ($\chi^2(6) = 560.49$, $p < .001$), as were Tone ($\chi^2(8) = 303.96$, $p < .001$) and Syllable ($\chi^2(6) = 3170.70$, $p < .001$). Significant two-way interactions were found between Tune and Tone ($\chi^2(4) = 171.16$, $p < .001$), Tune and Syllable ($\chi^2(3) = 352.55$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 193.80$, $p < .001$). In addition, the three-way interaction among Tune, Tone, and Syllable was significant ($\chi^2(2) = 69.80$, $p < .001$). The post-hoc analysis showed a pattern in which the first syllable of the question tune was generally shorter than was the second syllable. This difference was statistically significant for T2T2 ($\beta = -9.97$, $t(2836.02) = -2.63$, $p = .009$) and T4T4 ($\beta = -12.18$, $t(2836.02) = -3.21$, $p = .001$), although it was not significant for T1T1 ($\beta = -4.22$, $t(2836.02) = -1.11$, $p = .266$). Furthermore, the second syllable of the question tune was significantly longer than was that of the statement tune across all three tones (T1T1: $= 84.74$, $t(2836.02) = 22.34$, $p < .001$; T2T2: $\beta = 16.72$, $t(2836.02) = 4.41$, $p < .001$; T4T4: $\beta = 34.14$, $t(2836.02) = 9.00$, $p < .001$). These findings suggest a lengthening effect in the second syllable of the question tune.

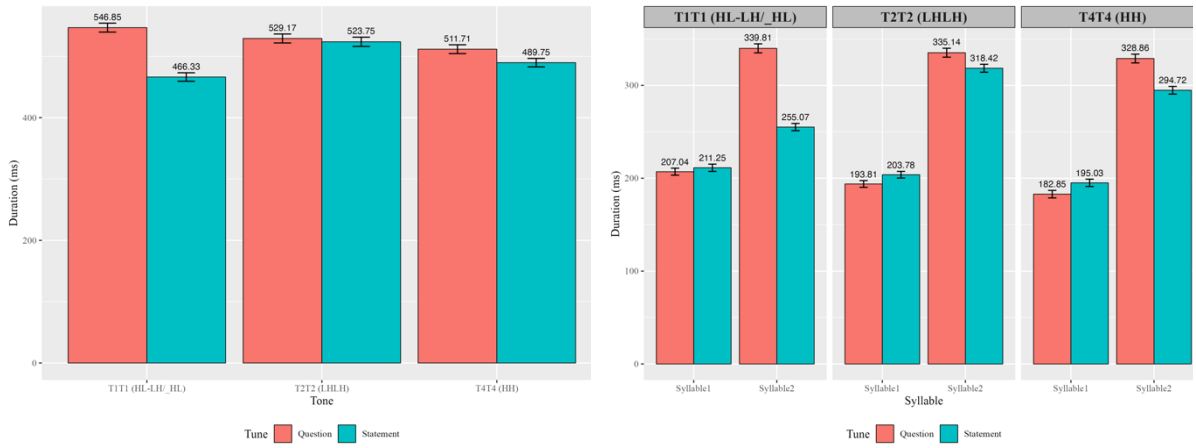


Figure 3.6 Duration (ms) at the whole-phrase level (left panel) and at the syllable level (right panel) across three tones

3.4.3 Mean Intensity

At the whole-phrase level (Table 3.8), significant main effects were observed for Tune, Tone, and their interactions on mean intensity (Tune: $\chi^2(3) = 501.33$, $p < .001$; Tone: $\chi^2(4) = 16.19$, $p = .003$; Tune x Tone: $\chi^2(2) = 6.59$, $p = .037$). Post-hoc tests further showed that questions consistently had a significantly higher mean intensity compared to statements across all the tones: T1T1 ($\beta = 3.20$, $t(1390) = 14.87$, $p < .001$), T2T2 ($\beta = 3.31$, $t(1390) = 15.39$, $p < .001$), and T4T4 ($\beta = 2.58$, $t(1390) = 12.02$, $p < .001$).

Table 3.8 Mean Intensity (dB) with SD in parentheses for three tones

Tone	Tune	Mean Intensity
T1T1	Q	66.13 (4.37)
	S	62.93 (3.96)
T2T2	Q	65.71 (4.40)
	S	62.40 (4.07)
T4T4	Q	65.89 (4.49)
	S	63.31 (3.81)

At the syllable level (Table 3.9), significant main effects were observed for Tune ($\chi^2(6) = 813.32$, $p < .001$), Tone ($\chi^2(8) = 55.24$, $p < .001$), and Syllable ($\chi^2(6) = 86.18$, $p < .001$), indicating that each of these factors independently impacted mean intensity. Significant two-way interactions were also found between Tune and Tone ($\chi^2(4) = 24.32$, $p < .001$), Tune and Syllable ($\chi^2(3) = 51.35$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 29.88$, $p < .001$). In addition, a significant three-way interaction among Tune, Tone, and Syllable ($\chi^2(2) = 6.21$, $p = .045$) was observed. Post-hoc analyses discovered that the mean intensity was consistently higher for questions than it was for statements across both syllables in all the tones. For the first syllable, questions had significantly higher intensity compared to statements in T1T1 ($\beta = 2.26$, $t(2836.02) = 9.44$, $p < .001$), T2T2 ($\beta = 2.61$, $t(2836.02) = 10.95$, $p < .001$), and T4T4 ($\beta = 1.78$, $t(2836.02) = 7.46$, $p < .001$). Similarly, for the second syllable, questions showed significantly greater intensity compared to statements for T1T1 ($\beta = 4.25$, $t(2836.02) = 17.81$, $p < .001$), T2T2 ($\beta = 3.53$, $t(2836.02) = 14.79$, $p < .001$), and T4T4 ($\beta = 2.81$, $t(2836.02) = 11.77$, $p < .001$).

Table 3.9 Mean Intensity (dB) with SD in parentheses by tone, syllable and tune

Tone	Sy.	Tune	Mean Intensity
T1T1	Sy.1	Q	65.54 (4.16)
		S	63.29 (3.82)
	Sy.2	Q	66.22 (4.69)
		S	61.97 (4.69)
T2T2	Sy.1	Q	64.86 (4.48)
		S	62.24 (4.06)
	Sy.2	Q	65.87 (4.55)
		S	62.34 (4.20)
T4T4	Sy.1	Q	64.83 (4.72)
		S	63.05 (3.91)
	Sy.2	Q	66.13 (4.60)
		S	63.32 (3.90)

3.4.4 Mean F0

At the whole-phrase level (Table 3.10), significant main effects were found for Tune ($\chi^2(3) = 982.34$, $p < .001$) and Tone ($\chi^2(4) = 80.78$, $p < .001$), while their interaction was not statistically significant (Tune x Tone: $\chi^2(2) = 2.39$, $p = .303$). Post-hoc pairwise comparisons of the mean F0 at this level indicated that, across all the tones, questions consistently showed significantly higher mean F0s compared to statements ((T1T1: $\beta = 56.11$, $t(1393.72) = 20.38$, $p < .001$; T2T2: $\beta = 61.34$, $t(1393.72) = 22.27$, $p < .001$; T4T4: $\beta = 61.30$, $t(1393.72) = 22.26$, $p < .001$).

Table 3.10 Mean F0 (Hz) with SD in parentheses for three tones

Tone	Tune	Mean F0
T1T1	Q	220.53 (82.25)
	S	164.42 (57.91)
T2T2	Q	226.70 (77.23)
	S	165.36 (54.16)
T4T4	Q	248.98 (89.81)
	S	187.69 (59.42)

At the syllable level (Table 3.11), significant main effects were found for Tune, Tone, and Syllable (Tune: $\chi^2(6) = 1769.70$, $p < .001$; Tone: $\chi^2(8) = 119.12$, $p < .001$; Syllable: $\chi^2(6) = 186.73$, $p < .001$). Significant two-way interactions were also observed between Tune and Syllable ($\chi^2(3) = 69.34$, $p < .001$) and between Tone and Syllable ($\chi^2(4) = 39.19$, $p < .001$). However, the interaction between Tune and Tone was not significant ($\chi^2(4) = 4.81$, $p = .308$), indicating that the effect of Tune on the mean F0 did not differ significantly across Tone levels. Similarly, the three-way interaction among Tune, Tone, and Syllable was not significant ($\chi^2(2) = 3.93$, $p = .140$). In addition, significant differences in the mean F0 were found between questions and statements across all the tones and for both syllables, with questions consistently having a higher pitch. Specifically, for T1T1, the first syllable of the question had a significantly higher pitch compared to the corresponding syllable in the statement ($\beta = 47.21$,

$t(2835.03) = 16.59$, $p < .001$), while the second syllable showed an even larger difference ($\beta = 64.78$, $t(2835.12) = 22.74$, $p < .001$). A similar pattern was found in T2T2: The first syllable displayed a mean F0 difference of 43.35 Hz ($t(2835.03) = 15.23$, $p < .001$), while the second syllable exhibited an even greater difference ($\beta = 68.42$, $t(2835.03) = 24.04$, $p < .001$). For T4T4, the first syllable differed between question and statement intonation by 51.23 Hz ($t(2835.03) = 18.00$, $p < .001$), with the second syllable showing an even larger pitch difference of 65.26 Hz ($t(2835.03) = 22.93$, $p < .001$).

Table 3.11 Mean F0 (Hz) with SD in parentheses by tone, syllable and tune

Tone	Sy.	Tune	Mean F0
T1T1	Sy.1	Q	213.21 (76.75)
		S	165.99 (56.11)
	Sy.2	Q	225.56 (89.80)
		S	160.35 (63.31)
T2T2	Sy.1	Q	203.61 (68.68)
		S	160.25 (52.34)
	Sy.2	Q	236.28 (81.38)
		S	167.86 (55.55)
T4T4	Sy.1	Q	237.60 (83.56)
		S	186.37 (58.96)
	Sy.2	Q	253.41 (93.40)
		S	188.16 (59.74)

3.4.5 F0 Extremes

At the whole-phrase level, for maximum F0, the main effects of Tune, Tone, and their interaction were significant (Tune: $\chi^2(3) = 718.87$, $p < .001$; Tone: $\chi^2(4) = 25.33$, $p < .001$; Tune x Tone: $\chi^2(2) = 16.12$, $p < .001$). Post-hoc pairwise comparisons of maximum F0 at the whole-phrase level showed that questions consistently exhibited significantly higher maximum F0 compared to statements across all the tones (T1T1: $\beta = 56.77$, $t(1393.72) = 14.86$, $p < .001$; T2T2: $\beta = 78.30$, $t(1393.72) = 20.49$, $p < .001$; T4T4: $\beta = 65.06$, $t(1393.72) = 17.03$, $p < .001$).

At the syllable level of maximum F0, Tune, Tone, and Syllable each showed significant main effects. Tune had a particularly strong influence ($\chi^2(6) = 1565.90$, $p < .001$), while Tone ($\chi^2(8) = 63.43$, $p < .001$) and Syllable ($\chi^2(6) = 87.99$, $p < .001$) also demonstrated considerable effects, suggesting that all three factors independently affected maximum F0. Significant two-way interactions were also observed, specifically between Tune and Tone ($\chi^2(4) = 12.49$, $p = .014$), Tune and Syllable ($\chi^2(3) = 31.50$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 41.81$, $p < .001$). Furthermore, a significant three-way interaction among Tune, Tone, and Syllable ($\chi^2(2) = 7.86$, $p = .020$) was found. The post-hoc analysis of maximum F0 confirmed that there were consistent differences between questions and statements across all the tones and syllables, with questions showing higher maximum F0. For example, the first syllable in questions in T1T1 had a significantly higher maximum F0 compared to statements ($\beta = 58.63$, $t(2836.03) = 16.99$, $p < .001$), while the difference was even larger in the second syllable ($\beta = 63.40$, $t(2836.03) = 18.37$, $p < .001$). Such trend persisted in T2T2, in which the first syllable differed by 56.02 Hz ($t(2836.03) = 16.23$, $p < .001$) and the second syllable showed an even larger increase of 80.00 Hz ($t(2836.03) = 23.18$, $p < .001$). For T4T4, questions exhibited a mean F0 increase of 56.12 Hz in the first syllable ($t(2836.03) = 16.26$, $p < .001$) and a 68.56 Hz rise in the second syllable ($t(2836.03) = 19.86$, $p < .001$).

Similarly, at the whole-phrase level for minimum F0, the main effects of Tune, Tone, and their interaction were also significant (Tune: $\chi^2(3) = 666.42$, $p < .001$; Tone: $\chi^2(4) = 166.03$, $p < .001$; Tune x Tone: $\chi^2(2) = 40.21$, $p < .001$). Post-hoc pairwise comparisons of minimum F0 at the whole-phrase level revealed that questions consistently exhibited significantly higher minimum F0 compared to statements across all the tones (T1T1: $\beta = 44.64$, $t(1390) = 19.96$, p

< .001; T2T2: $\beta = 25.39$, $t(1390) = 11.35$, $p < .001$; T4T4: $\beta = 40.25$, $t(1390) = 18.00$, $p < .001$). See Table 3.12 for the average maximum and minimum F0 at the whole-phrase level.

Table 3.12 Maximum and minimum F0 (Hz) with SD in parentheses for three tones

Tone	Tune	Maximum F0	Minimum F0
T1T1	Q	253.21 (102.33)	170.10 (55.32)
	S	196.54 (79.47)	125.46 (45.95)
T2T2	Q	267.01 (92.23)	164.58 (52.25)
	S	188.71 (64.05)	139.20 (45.19)
T4T4	Q	265.60 (94.32)	210.29 (68.10)
	S	200.54 (63.26)	170.03 (53.40)

At the syllable level for minimum F0, all main effects and interactions were significant. The analysis revealed strong main effects for Tune ($\chi^2(6) = 1552.70$, $p < .001$), Tone ($\chi^2(8) = 200.43$, $p < .001$), and Syllable ($\chi^2(6) = 505.92$, $p < .001$). Significant interactions were found between each *p*air of factors: Tune and Tone ($\chi^2(4) = 42.08$, $p < .001$), Tune and Syllable ($\chi^2(3) = 236.38$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 64.24$, $p < .001$), as well as the three-way interaction ($\chi^2(2) = 21.58$, $p < .001$). The post-hoc analysis of minimum F0 revealed that there were consistent differences between questions and statements across all the tones and for both syllables, with questions showing higher minimum F0. The first syllable of the question in T1T1 had a significantly higher minimum F0 compared to the corresponding syllable in the statement ($\beta = 28.65$, $t(2836.02) = 10.69$, $p < .001$), while the difference was even greater in the second syllable ($\beta = 73.83$, $t(2836.02) = 27.54$, $p < .001$). A similar trend appeared in T2T2, in which the first syllable differed by 23.12 Hz ($t(2836.02) = 8.62$, $p < .001$), and the second syllable showed a larger increase of 55.97 Hz ($t(2836.02) = 20.88$, $p < .001$). The minimum F0 in the first syllable in T4T4 was higher by 38.22 Hz ($t(2836.02) = 14.26$, $p < .001$), while the second syllable showed a difference of 58.49 Hz ($t(2836.02) = 21.82$, $p < .001$). Both higher

maximum and minimum F0s suggests that the question pitch is lifted. See Table 3.13 for the average maximum and minimum F0s at the syllable level.

Table 3.13 Maximum and minimum F0 (Hz) with SD in parentheses by tone, syllable and tune

Tone	Sy.	Tune	Maximum F0	Minimum F0
T1T1	Sy.1	Q	246.89 (96.19)	173.27 (56.10)
		S	188.26 (66.29)	144.62 (46.98)
	Sy.2	Q	249.20 (102.24)	205.06 (78.45)
		S	185.81 (79.81)	131.23 (53.87)
T2T2	Sy.1	Q	235.88 (84.51)	164.68 (52.29)
		S	179.86 (60.58)	141.56 (45.43)
	Sy.2	Q	266.26 (92.44)	207.90 (71.30)
		S	186.26 (63.63)	151.92 (51.17)
T4T4	Sy.1	Q	253.67 (90.77)	211.63 (68.29)
		S	197.54 (62.46)	173.41 (53.69)
	Sy.2	Q	262.01 (94.57)	239.05 (88.87)
		S	193.44 (61.71)	180.57 (57.72)

3.4.6 F0 Range

At the whole-phrase level (on the left of Figure 3.7), significant main effects were observed for Tune, Tone, and their interaction on F0 range (Tune: $\chi^2(3) = 288.60$, $p < .001$; Tone: $\chi^2(4) = 163.13$, $p < .001$; Tune x Tone: $\chi^2(2) = 76.17$, $p < .001$). Questions consistently displayed a significantly higher F0 range compared to statements across all the tones. For T1T1, the F0 range difference was estimated at 12.12 Hz ($\beta = 12.12$, $t(1390.79) = 3.63$, $p = .0003$), while the difference for T2T2 was significantly larger at 52.91 Hz ($\beta = 52.91$, $t(1390.79) = 15.83$, $p < .001$). In T4T4, the F0 range difference was 24.80 Hz ($\beta = 24.80$, $t(1390.79) = 7.42$, $p < .001$), further demonstrating the consistent trend across the tones.

At the syllable level (on the right of Figure 3.7), the main effects for Tune ($\chi^2(6) = 398.7$, $p < .001$), Tone ($\chi^2(8) = 199.91$, $p < .001$), and Syllable ($\chi^2(6) = 221.19$, $p < .001$) were significant, as were all the two-way interactions, Tune and Tone interacted significantly ($\chi^2(4) = 99.05$, $p < .001$), as did Tune and Syllable ($\chi^2(3) = 120.29$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 54.51$, $p < .001$). A significant three-way interaction among Tune, Tone, and Syllable ($\chi^2(2) = 47.12$, $p < .001$) was also found. In the post-hoc analysis, the first syllable of the question in T1T1 had a significantly larger F0 range compared to the corresponding syllable in the statement ($\beta = 29.98$, $t(2836) = 11.14$, $p < .001$). By contrast, the second syllable displayed a significantly smaller F0 range compared to the statement ($\beta = -10.44$, $t(2836) = -3.88$, $p = .0001$), suggesting a more obvious fall in the second syllable of T1 in the statement tune. For T2T2, both syllables in questions had a greater F0 range than they did in statements, with a larger difference in the first syllable ($\beta = 32.90$, $t(2836) = 12.23$, $p < .001$) compared to the second syllable ($\beta = 24.02$, $t(2836) = 8.93$, $p < .001$). Similarly, for T4T4, the first syllable in questions showed a higher F0 range ($\beta = 17.90$, $t(2836) = 6.66$, $p < .001$), while the second syllable exhibited a smaller but still significant increase ($\beta = 10.07$, $t(2836) = 3.75$, $p = .0002$).

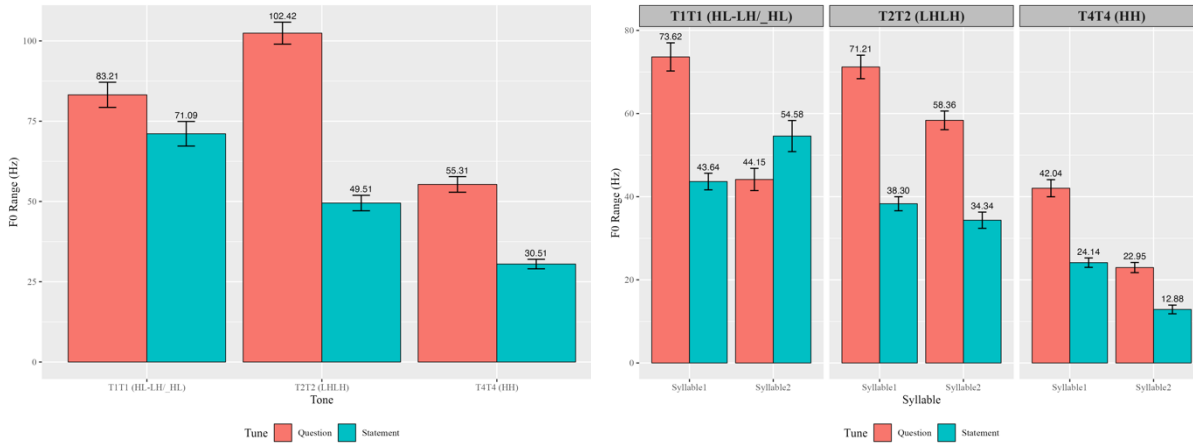


Figure 3.7 F0 range (Hz) at the whole-phrase level (left panel) and at the syllable level (right panel) across three tones

3.4.7 Maximum F0 and Minimum F0 Alignment

The maximum and minimum F0 alignments were calculated based on the timing of the maximum or minimum F0 relative to the syllable length and to the whole-phrase length. For the maximum F0 alignment, the main effects of Tune, Tone, and their interaction were significant at the whole-phrase level (Tune: $\chi^2(3) = 188.79$, $p < .001$; Tone: $\chi^2(4) = 140.79$, $p < .001$; Tune x Tone: $\chi^2(2) = 44.17$, $p < .001$). The post-hoc analysis showed significant differences in the maximum F0 alignment between questions and statements for T2T2 ($\beta = 0.21$, $t(1389.94) = 9.13$, $p < .001$) and T4T4 ($\beta = 0.25$, $t(1389.94) = 10.73$, $p < .001$), suggesting that the peak pitch in questions occurred relatively later than in statements for these tones. However, the difference was not statistically significant for T1T1 ($\beta = 0.04$, $t(1389.94) = 1.85$, $p = .065$).

At the syllable level, there were significant main effects for Tune, Tone, and Syllable (Tune: $\chi^2(6) = 256.19$, $p < .001$; Tone: $\chi^2(8) = 473.53$, $p < .001$; Syllable: $\chi^2(6) = 871.15$, $p < .001$). Significant two-way interactions were also observed between Tune and Tone ($\chi^2(4) = 25.57$, $p < .001$), Tune and Syllable ($\chi^2(3) = 52.64$, $p < .001$), and Tone and Syllable ($\chi^2(4) = 363.63$, $p < .001$), as well a significant three-way interaction among Tune, Tone, and Syllable ($\chi^2(2) = 7.93$, $p = .019$). The post-hoc analysis showed that, in T1T1, the first syllable in questions showed a significantly later maximum F0 location compared to statements ($\beta = 0.22$, $t(2835.97) = 7.69$, $p < .001$), whereas no significant difference was found for the second syllable ($\beta = 0.00$, $t(2835.97) = -0.14$, $p = .887$). For T2T2 and T4T4, the first syllable in questions also showed a later maximum F0 than it did in statements (T2T2: $\beta = 0.19$, $t(2835.97) = 6.63$, $p < .001$; T4T4: $\beta = 0.32$, $t(2835.97) = 11.08$, $p < .001$). Similarly, the second syllable exhibited a later

maximum F0 in questions, although with smaller effect sizes (T2T2: β = 0.12, t(2835.97) = 4.27, p < .001; T4T4: β = 0.14, t(2835.97) = 4.83, p < .001).

For the minimum F0 alignment at the whole-phrase level, significant main effects were found for Tune, Tone, and their two-way interaction (Tune: χ^2(3) = 392.33, p < .001; Tone: χ^2(4) = 152.09, p < .001; Tune x Tone: χ^2(2) = 77.99, p < .001). The post-hoc result indicated that, questions consistently exhibited an earlier minimum F0 location compared to statements across all tones at the whole-phrase level. Specifically, for T1T1, the minimum F0 location was significantly earlier in questions (β = -0.48, t(1389.94) = -18.44, p < .001), and similar trends were observed for T2T2 (β = -0.20, t(1389.94) = -7.70, p < .001) and T4T4 (β = -0.19, t(1389.94) = -7.29, p < .001).

At the syllable level, similar findings were found that significant main effects and interactions were observed across Tune, Tone, and Syllable, as well as in their two-way and three-way interactions (Tune: χ^2(6) = 215.68, p < .001; Tone: χ^2(8) = 265.18, p < .001; Syllable: χ^2(6) = 1024.5, p < .001; Tune x Tone: χ^2(4) = 55.75, p < .001; Tune x Syllable: χ^2(3) = 68.72, p < .001; Tone x Syllable: χ^2(4) = 188.17, p < .001; Tune x Tone x Syllable: χ^2(2) = 49.12, p < .001). The post-hoc result showed that, for T1T1, the first syllable in questions had a significantly earlier minimum F0 location than it did in statements (β = -0.26, t(2835.97) = -10.54, p < .001), while no significant difference was observed for the second syllable (β = 0.03, t(2835.97) = 1.12, p = .263). For both T2T2 and T4T4, questions displayed a significantly earlier minimum F0 location compared to statements for both syllables; in the first syllable (T2T2: β = -0.10, t(2835.97) = -4.17, p < .001; T4T4: β = -0.14, t(2835.97) = -5.62, p < .001) and, in the second syllable (T2T2: β = -0.08, t(2835.97) = -3.42, p = .0006; T4T4: β = -0.17, t(2835.97) = -7.09,

$p < .001$). See Table 3.14 and Table 3.15 for the maximum F0 and minimum F0 alignments at the whole-phrase level and at the syllable level, respectively.

Table 3.14 Maximum F0 and minimum F0 alignment with SD in parentheses for three tones

Tone	Tune	Maximum F0 Alignment	Minimum F0 Alignment
T1T1	Q	37.12 (16.53)	22.18 (35.88)
	S	32.77 (24.85)	70.33 (35.29)
T2T2	Q	83.32 (21.50)	4.65 (8.18)
	S	61.86 (37.65)	24.75 (28.14)
T4T4	Q	60.78 (28.85)	13.98 (28.78)
	S	35.56 (33.86)	33.00 (40.41)

Table 3.15 Maximum F0 and minimum F0 alignment with SD in parentheses by tone, syllable and tune

Tone	Sy.	Tune	Maximum F0 Alignment	Minimum F0 Alignment
T1T1	Sy.1	Q	82.94 (30.05)	15.81 (23.26)
		S	60.67 (43.18)	41.45 (39.15)
	Sy.2	Q	10.76 (20.16)	75.78 (26.71)
		S	11.17 (26.80)	73.06 (23.52)
T2T2	Sy.1	Q	89.45 (21.58)	11.54 (11.89)
		S	70.24 (40.43)	21.69 (17.28)
	Sy.2	Q	77.41 (29.00)	29.05 (19.15)
		S	65.05 (39.84)	37.38 (13.40)
T4T4	Sy.1	Q	77.12 (34.83)	13.75 (21.24)
		S	45.05 (42.09)	27.43 (32.56)
	Sy.2	Q	51.73 (31.37)	44.55 (46.10)
		S	37.75 (30.33)	61.80 (40.31)

3.5 Discussion

3.5.1 Statistical Results Revisited

The statistical results from the previous section are summarised in Table 3.16.

Table 3.16 Summary of the statistical results of Production Study 1

Factor	Level	Summary
F0 Change Rate	Whole-Phrase	• Questions had significantly higher slopes and intercepts, indicating a higher F0 change rate compared to statements. • T1T1: Questions had a positive slope while statements had a negative slope. • T2T2 and T4T4: Both questions and statements had positive slopes.
	Syllable	• Intercept: Questions had significantly larger intercepts for both syllables compared to statements. • Slopes: o First syllable: Both tunes had positive slopes, indicating a rising tendency. o Questions had significantly steeper slopes, indicating a stronger rising tendency. o Second syllable: ▪ T1T1: Slopes were negative for both tunes, but questions were less negative than statements, indicating less of a falling tendency in questions. ▪ T2T2: Positive slopes for both tunes, with questions being significantly steeper. ▪ T4T4: Positive slope for questions and negative slope for statements, with questions showing a larger slope.
Duration	Whole-Phrase	• Questions were longer than statements overall; significantly longer for T1T1 and T4T4, but not significantly for T2T2. T1T1 difference was the largest.
	Syllable	• First syllable: Questions were shorter than statements; significantly for T2T2 and T4T4, but not significantly for T1T1. • Second syllable: Questions were significantly longer than statements. T1T1 had the largest difference, followed by T4T4 and then T2T2.
Mean Intensity	Whole-Phrase	• Questions had a significantly higher mean intensity than statements.
	Syllable	• Questions had a significantly higher mean intensity for both syllables than statements.
F0 (Mean, Maximum, Minimum)	Whole-Phrase	• Questions had significantly higher mean, maximum, and minimum F0 than statements.
	Syllable	• Questions had significantly higher mean, maximum, minimum F0 for both syllables than statements.

		• The second syllable had larger differences between questions and statements compared to the first syllable.
F0 Range	Whole-Phrase	• Questions had a significantly larger F0 range than statements. • The largest difference was found in T2T2, followed by T4T4 and T1T1.
	Syllable	• T1T1: o First syllable: The F0 range was significantly larger in questions than in statements. o Second syllable: The F0 range was significantly smaller in questions than in statements. • T2T2 and T4T4: o Questions had a significantly larger F0 range than statements. o The first syllable showed a greater difference between questions and statements than the second syllable.
Maximum F0 Alignment	Whole-Phrase	• Maximum F0 occurred later in questions than in statements; significantly for T2T2 and T4T4, but not significantly for T1T1.
	Syllable	• T1T1: o First syllable: Maximum F0 occurred significantly later in questions than in statements. o Second syllable: No significant differences. • T2T2 and T4T4: • Both syllables had later maximum F0 in questions than in statements, with the second syllable showing a smaller difference between questions and statements.
Minimum F0 Alignment	Whole-Phrase	• Minimum F0 occurred significantly earlier in questions than in statements.
	Syllable	• T1T1: o First syllable: Minimum F0 occurred significantly earlier in questions than in statements. o Second syllable: No significant differences. • T2T2 and T4T4: o Both syllables had significantly earlier minimum F0 alignment in questions than in statements.

It is observed that the acoustic realisation of yes-no questions in GuanM is encoded through multiple parallel acoustic features. The following sections discuss the temporal organisation, pitch trajectory modifications, and global prosodic settings.

3.5.2 Duration

Duration acts a critical temporal cue for interrogativity in GuanM, varying systematically according to the compatibility between the lexical tone and the interrogative contour. At the whole-phrase level, questions were generally longer than their declarative counterparts, with a significant rise in duration observed for the T1T1 (rising-falling) and T4T4 (high-level) sequences. Nevertheless, the T2T2 (rising-rising) sequence did not exhibit the significant lengthening, which can be attributed to the articulatory compatibility between the lexical tone and intonation. Since T2 is inherently rising, the superimposition of a high boundary tone (H%) creates a facilitative interaction, meaning the intonational rise can be achieved without requiring additional time. In contrast, for T1T1 and T4T4, the interaction is incompatible. For T1T1 (rising-falling), the speaker must counteract the downward trajectory to accommodate the H%; for T4 (high-level), the speaker must generate a rise where none exists lexically.

At the syllable level, an asymmetric distribution of duration was found. Speakers consistently compressed the initial syllable whilst significantly lengthening the final syllable in questions. This suggests a strategy of temporal reorganisation to facilitate the realisation of the boundary tone. The final syllable acts as the primary locus for the intonational contrast, requiring extended duration to fully realise the H%. This effect is most evident in the second syllable of T1T1. Crucially, although the T1 contour remains phonetically falling in questions, the presence of the H% boundary tone demands a significantly higher ending point compared to the statement counterparts. Consequently, the speaker must exert articulatory effort to mitigate the steepness of the fall, which acts as a slowing mechanism: the speaker requires more time to execute a controlled, shallower fall that sustains the pitch height at the offset in questions, as opposed to the rapid, unrestricted drop observed in statements. These findings align with previous research on SC, for instance, Yuan (2004) observed that final tones with conflicting

trajectories (such as low-rising T3 and falling T4) were significantly lengthened in questions, whereas high (T1) and rising (T2) tones, which are more compatible with high registers, showed less variation.

3.5.3 F0 Change Rate

The F0 change rate was quantified using the intercept and slope of a fitted line, which showed evidence for the separation of global and local prosodic mechanisms. First, the intercept values serve as the primary indicator of the global register effect. Across all tone sequences and both syllables, questions exhibited significantly higher intercept values than statements, indicating that the F0 starting point is systematically raised regardless of the lexical tone, suggesting that speakers employ a global register expansion strategy to signal interrogativity from the very onset of the utterance.

In contrast, the slope analysis reveals that interrogativity is not merely a global lift, but involves a distinct H% at the phrase edge, which is manifested by the specific pitch trajectories on the final syllable, which varies according to the lexical environment.

For T1T1 (rising-falling), while the first syllable maintained its positive slope (likely preserving the tone dissimilation rule), the second syllable revealed a clear interaction between the falling lexical target and the rising intonation. Statements showed a steep negative slope, whereas questions showed a significantly less negative slope, indicating a mitigated fall in which the H% exerts an upward pull, counteracting the lexical declination even if it does not fully reverse it.

Regarding T2T2 (rising-rising), the H% operates collaboratively with the lexical tone. Questions had a significantly steeper positive slope than statements on the final syllable. While statements showed a minimal positive tendency (likely dampened by a low boundary tone or declination), the question intonation accelerated the rising pattern, confirming that the H% target actively enhances the lexical rise.

In the case of T4T4 (high-level), this sequence offers the clearest evidence of an opposite change. While statements displayed a negligible negative slope on the final syllable (possibly indicative of natural declination in statements), questions exhibited a positive slope. This qualitative slope inversion, turning a flat or nearly-fall contour into a rise, cannot be explained by global register raising alone. It confirms the presence of a specific, phonologically encoded H% at the end of the phrase edge. This H% will be elaborated on further in Section 3.5.6.

3.5.4 Register and Intensity

The first question concerns whether syntactically unmarked yes-no questions in GuanM exhibit a declination or a rising tendency. The F0 change rate analysis demonstrated that all tonal sequences displayed positive slopes at the whole-utterance level, confirming that the global intonation pattern of questions in GuanM is characterised by a consistent upward trajectory, effectively blocking the natural declination typically observed in statements.

Furthermore, questions consistently exhibited higher mean, maximum, and minimum F0 values compared to statements at both the whole-phrase and syllable levels, indicating that questions are produced with a globally raised pitch register, where both the F0 floor and ceiling are elevated. Such global register rising is also reported by F. Liu and Xu (2005) for SC, who found that the intonation pattern of yes-no questions was characterised by pitch raising that

followed an exponential (or even double-exponential) function across the entire sentence, rather than a simple final pitch rise. Additionally, questions were produced with higher intensity across both syllables, suggesting that the physiological mechanism involves increased subglottal pressure to support this global raising.

However, it is crucial to separate this global register effect from the local phonological impact at the phrase edge. While the global raising elevates the entire contour, the F0 difference between questions and statements was not uniform. The second syllable (also the last syllable in this study) displayed a significantly larger pitch difference than the first, indicating that the pitch rise is specifically intensified at the endpoint. This cumulative widening of the pitch gap demonstrates that the final F0 height is not merely the result of a static register lift, but of a superposition of two effects:

1. Global Register Raising: Provides an elevated baseline across the entire utterance.
2. Local Phonological Target: Exerts an additional upward pull at the endpoint.

This observation is consistent with P. Jiang and Chen (2011), who reported global rising intonation and prominence at syllable-final positions in yes-no questions in SC.

3.5.5 The Mechanism of the Final Rise

In this study, we posit that syntactically unmarked yes-no questions in GuanM are characterised by a phonological final rise. However, the surface realisation of this rise is not uniform; rather, it is tone-dependent, influenced by the interaction between pitch range expansion, global register raising, and temporal realignment at the phrase edge.

First, the analysis of F0 range reveals how the final rise interacts with different lexical tones. Overall, questions exhibited a significantly larger F0 range than statements, particularly on the

final syllable. Substantial expansion was observed in the T2T2 (rising) and T4T4 (high) sequences, suggesting that tones with inherently rising or high properties were further exaggerated by the question intonation. However, the T1T1 sequence displayed a conflicting pattern. While the first syllable expanded, the second syllable, an inherently falling tone, showed a reduced F0 range compared to statements, which is interpreted as a phonetic compression at the phrase-final position. To accommodate the final rise without destroying the essential falling nature of T1, the speaker creates a mitigated fall, compressing the range to prevent the pitch from dropping as low as it typically would in a statement.

In addition to F0 range modification, as discussed in the previous section, the F0 scaling data indicates that interrogative utterances were produced with an overall raised pitch register, evidenced by higher maximum and minimum F0 values compared to statements. Moreover, the magnitude of this difference was not uniform across the phrase; the most significant difference in F0 extrema occurred on the second syllable, indicating the cumulative widening of the gap at the end of the utterance, which suggests that, while the register lift is global, the phrase-final position serves as the primary locus of interrogative marking.

Furthermore, the alignment of F0 minimum and maximum provides temporal evidence for this phrase-final target. Generally, the Maximum F0 occurred later in questions than in statements, particularly significantly for the T2T2 and T4T4 sequences. This phenomenon, often described as peak delay (Gussenhoven, 2004), suggests that the pitch trajectory is actively extended towards the right edge of the phrase to facilitate the realisation of the high boundary tone. Concurrently, the Minimum F0 occurred significantly earlier in questions, indicating that speakers reached their lowest pitch point significantly earlier than in statements by speeding up the initial downward pitch movement, essentially compressing it, so that speakers arrive at

the low pitch target more quickly. By doing this, they create more time for the upward pitch movement that follows, which is crucial for signalling that the utterance is a question.

The T1T1 sequence, however, presents an exception to this pattern of temporal reorganisation. While the first syllable of T1T1 exhibited the expected peak delay, the second syllable did not show a significant shift in peak alignment, which likely results from the specific interactional outcome between the falling lexical tone and the rising intonation. Since the falling movement is counteracted by the question's rising intonation pressure, the need for temporal adjustments is reduced. Instead, the interaction manifests as a phonological compromise: the second syllable retains its essential falling characteristic (preserving lexical identity), but the falling trajectory is less steep than in statements. Consequently, the final portion ends relatively high, contributing to the overall heightened register of questions. The levelling-out effect produced by the final rise means that pitch range compression serves to accommodate the intonational target while maintaining the tonal contour, rendering substantial temporal shifts unnecessary.

3.5.6 Boundary Tone

Given the preceding discussion regarding the mechanism of the final rise, this study proposes that syntactically unmarked yes-no questions in GuanM are phonologically marked by a High Boundary Tone (H%), which functions as a categorical tone associated with the right edge of the intonational phrase (IP), but its phonetic realisation is gradient and determined by its superimposition onto the final lexical tone. Consequently, the H% is an inherently rising tone, but its surface form is constrained by the lexical environment: it may surface as an enhanced rising contour, a rising contour, or a mitigated fall. Specifically, the H% exerts a consistent upward pull that modifies the slope and trajectory of the lexical tone relative to the statement baseline: rising tones are realised with a significantly greater rise than the inherent lexical

contour, while falling tones exhibit a mitigated fall that descends less steeply than the lexical falling tone, but the H% does not change the lexical tone identity, and level tones show the rising contour.

To contextualise this proposal, it is necessary to revisit the definition and function of boundary tones within the AM framework. The key concepts in the AM framework are the pitch accents that are related to prominence, and the boundary tones that mark the edges of phrases or domains (Ladd, 2008). Originally, Pierrehumbert (1980) distinguished between phrase accents (H-, L-) and boundary tones (H%, L%), defining the latter as single targets associated with the very edge of an IP. In Pierrehumbert's analysis of English, the H% boundary tone indicates a final rise, which is straightforward after a Low phrase accent, but requires an upstep rule after a high phrase accent to ensure the boundary tone is realised higher than the preceding high target, preventing a sustained level pitch. Conversely, the L% boundary tone indicates the absence of a final rise. After a Low phrase accent, it signals a gradual fall to the bottom of the speaking range; after a high phrase accent, it results in a sustained level pitch. Arvaniti and Baltazani (2005) employed three boundary tones: H%, L%, and !H% in Greek, where !H% represents a mid-pitch level, effectively resolving the ambiguity of the "sustained level pitch" by providing a specific target that does not require phrase accent triggers.

However, cross-linguistic evidence suggests that boundary specifications can be more complex than a single binary choice. Hayes and Lahiri (1991) demonstrated that boundary tones can consist of tonal sequences. They also claimed that the boundaries assumed with the intonational phonology literature are not different from the phrasal phonology assumptions. Thus, the % intonational boundary tone matches to the edge of an intonational phrase while the T- phrasal tone aligns to a phonological phrase. In Bengali, yes-no questions are marked by a HL

boundary tone, meaning that interrogativity does not necessarily require a final rise; instead, the pitch rises smoothly to the final syllable before falling rapidly at the boundary. This is exemplified in (12), with the corresponding pitch trajectory shown in Figure 3.8.

(12) tumi-ki kágoǰɔla-ke dek^{h}le?

you Q-part. newspaperman-obj. saw

"Did you see the newspaperman?" (yes/no question nucleus)

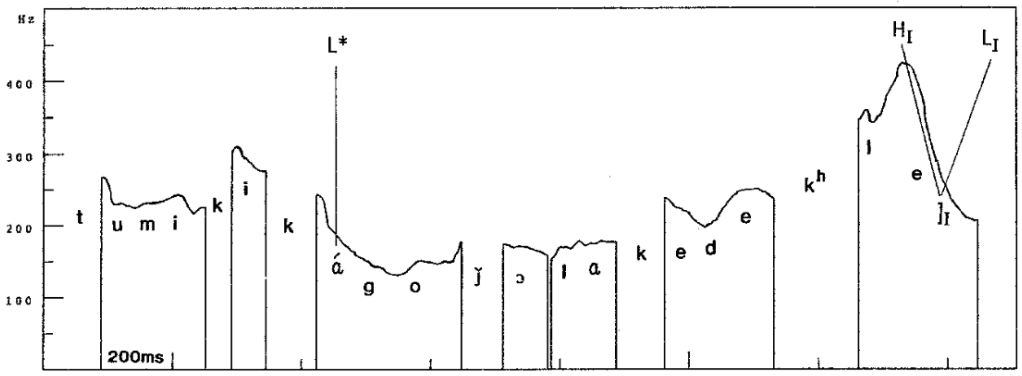


Figure 3.8 The pitch trajectory of yes-no question in Bengali (cited from Hayes & Lahiri, 1991)

In this contour (L* H$_I$ L$_I$), the pitch does not simply rise; instead, it rises smoothly from the low nucleus to the final syllable, whereupon it falls rapidly at the boundary. Such HL contours, aligned to the edge of an intonational phrase, are also the strategy of marking questions in African languages, as discussed in Section 3.1.

To further explain the mechanism of boundary tones in tonal contexts, it is necessary to look beyond Sinitic languages to African Bantu languages. Although these languages differ typologically, typically featuring register (level) tones rather than the contour tones characteristic of Chinese languages, they face an identical phonological challenge: how to

integrate intonational targets with lexical tonal contrasts. Hyman and Monaka (2011) propose a three-way typology for tone-intonation interaction: accommodation (co-existence, often via register change), submission (intonation overrides tone), and avoidance (intonation is minimised). Here we particularly examine the superimposition, exemplified in Embosi and Shingazidja.

In Embosi, a two-tone language (H and L) without downdrift, boundary tones are not simply inserted linearly into the tonal string but are superimposed upon it. Rialland and Aborobongui (2016) propose a Dual Register Model, where intonation functions by expanding the register above or below the lexical tone norms. This model involves a basic register for lexical tone realisations and enlargements above and below this basic register, triggered by extra-high or extra-low boundary tones. The H% in Embosi functions similarly to a continuation rise in languages such as English and appears in comparable contexts, particularly in juxtaposed assertive sentences. Importantly, the H% is not fixed rigidly to the final syllable; rather, it is attracted to the last H lexical tone of the phrase. When the final tone of the phrase is a lexical H, the H% superimposes onto it, creating an extra-high realisation with a rising contour. When the final tone is L, the H% seeks the preceding H tone to realise this register expansion, resulting in a falling final contour despite the extra-high peak occurring earlier in the phrase.

Such tone-intonation interaction becomes even more complicated in Embosi yes-no questions, which employ a HL% boundary contour (a combination of H% and L%) that creates a significant expansion of the pitch range compared to declarative utterances. The H% component targets the last lexical high tone, raising it to an extra-high realisation, while the L% associates with the final boundary and produces a strong lowering effect, particularly on final H tones. When the sentence ends with a L tone, the H% appears on the last H tone, making it

extra-high, as shown on the right side of Figure 3.9. However, if there is an H tone at the end of the utterance, this final H is lowered by the L% boundary tone, and the H% is pushed forward to a preceding H tone, extending the contour over a larger domain, as shown in Figure 3.10.

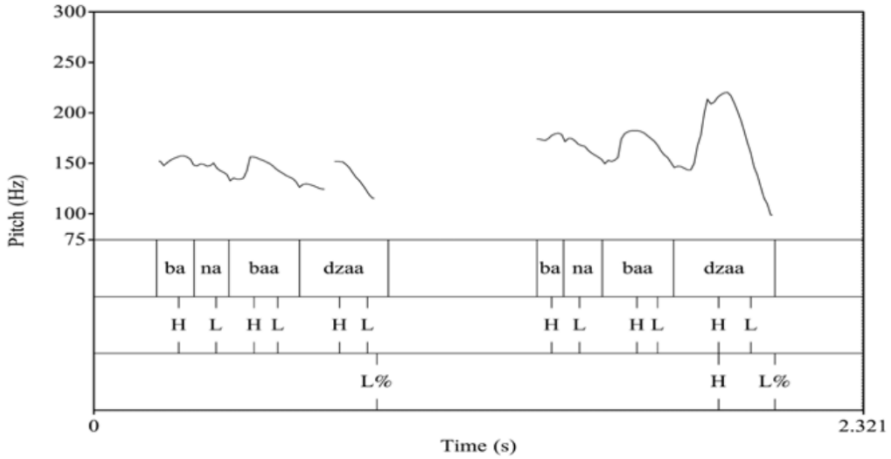


Figure 3.9 F0 contour of [bánabáadzáa].

Left panel: "The children eat." Right panel: "Do the children eat?" (Adapted from Rialland & Aborobongui, 2016)

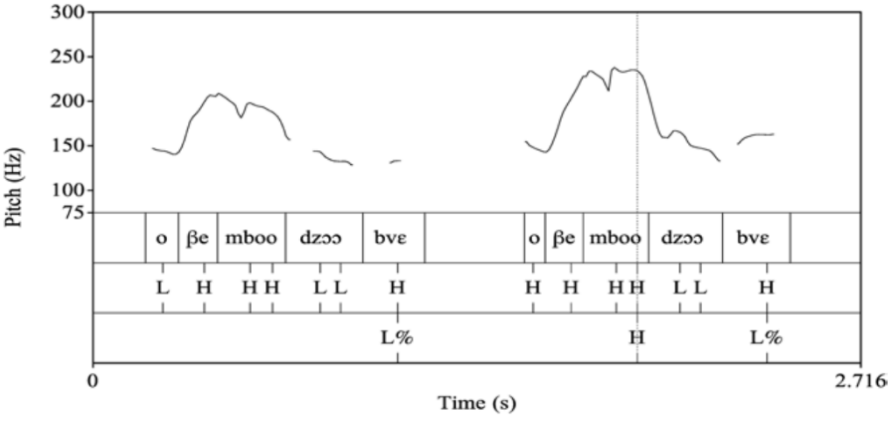


Figure 3.10 F0 contours of [oβémbóódzɔɔbvɛ́].

Left panel: "The journey was good." Right panel: "Was the journey good?" (Adapted from Rialland & Aborobongui, 2016)

Similarly, in Shingazidja (Bantu G44a), lexical tones can determine the physical placement of intonational targets. Patin (2016) observes that in polar questions, a super-high intonational tone (analysed as a LH* accent) typically targets the penultimate syllable. However, if the final syllable already carries a lexical high tone, the intonational target is blocked from its canonical position and is displaced (retracted) to the antepenult.

While African languages have shown examples of superimposition in level-tone systems, Sinitic languages offer the most relevant comparison for GuanM due to the shared presence of contour tones. Within the Sinitic family, there are two distinct strategies for marking interrogativity: the dominance model and the superimposition model. Cantonese provides a clear example of the former, where intonation overrides the lexical tone. J. K.-Y. Ma et al. (2006a) found that the H% in Cantonese yes-no questions replaces or dominates the underlying tonal contour. Specifically, regardless of whether the canonical tone is falling, level, or rising, all six tones exhibit a rising F0 trajectory at the sentence-final position in questions[10], as shown in Figure 3.11.

[10] B. R. Xu and Mok (2011) show that the rising is superimposed onto the latter half of the final syllable, while the first half of the tonal contour remains relatively intact. As a result, lower tones (T4, T5, T6) risk being misidentified as the high-rising tone (T2) during perception of questions since T2 shares the low start with those tones at question-final position (J. K.-Y. Ma et al., 2006a). Overall, lexical tones undergo significant convergence rather than complete neutralisation.

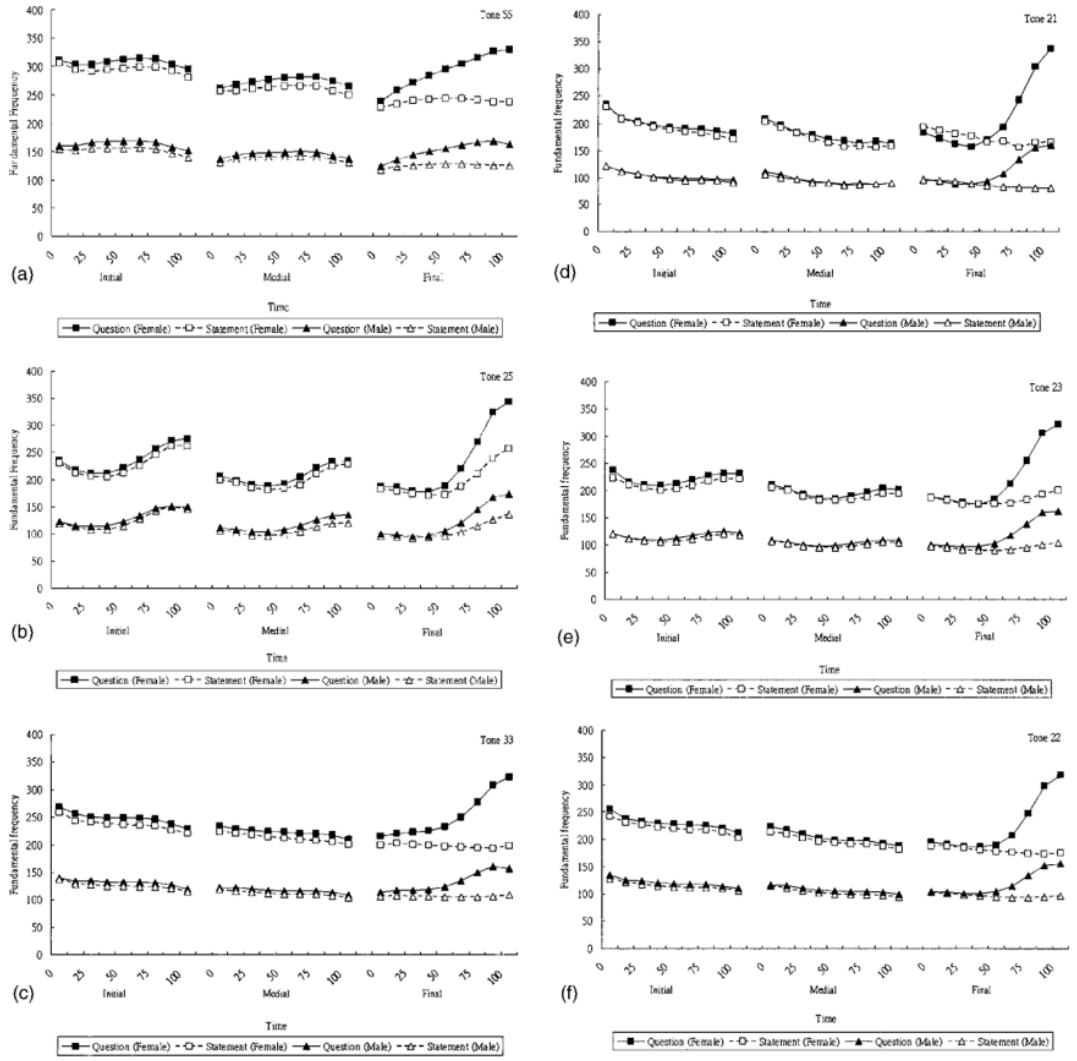


Figure 3.11 Yes-no questions pitch contours from six tones at the sentence-final positions in Cantonese (adapted from J. K.-Y. Ma et al., 2006a)

SC instead illustrates the superimposition relation, in which the shape of the lexical tone is largely preserved, meaning that the H% is not necessarily forced to be a final rise; rather, it modifies the pitch range and register of the existing lexical contour. The existence of the boundary tone concept in SC, distinct from global register, is well-established in the literature and formally codified in transcription systems. M. Lin (2004) argues that the primary marker of interrogativity in SC is a boundary tone located strictly at the phrase edge. His acoustic analysis revealed that the register of the endpoint (or the final F0 slope) was more significant

than the starting point in differentiating questions from statements. Specifically, for tones T1, T2, and T4, the F0 curve ended higher in questions, whereas for T3, it exhibited a rising pattern compared to the falling shape in statements. Lin concluded that intonation modifies tone in a hierarchical manner: tones define the F0 contour, while the boundary tone modifies the pitch register and range of that contour without erasing its shape. The Mandarin ToBI (M_ToBI) system, documented by Peng et al. (2005), explicitly separates the local edge effects from the global range effects. It assigns the tag H% to mark the final rising pitch at the right edge of yes-no questions (the local target), while simultaneously using a separate tag, %q-raise, to capture the overall pitch range expansion (the global macro cue). The system contains three tiers: the romanisation tier for underlying lexical tones, the sandhi tier for rhythm-related effects, and a third tier dedicated to boundary tones and pitch range variations. Unlike the Cantonese ToBI (C_ToBI) system where lexical tones and boundary tones are transcribed together on the same tier, the separation between lexical tone and boundary tone in the M_ToBI system is necessary because the lexical tone is not overridden; thus, the boundary tone must be modelled as an independent entity interacting with, rather than replacing, the lexical tone. Furthermore, Jiang and Chen's (2011) proposal of dual-edge marking ("H% ... H%") in SC questions, also fundamentally relies on the existence of the H% boundary tone to explain higher pitch targets in the question intonation, differing only in its placement at both left and right edge of the IP.

The phonological framework of GuanM mirrors this structure of the question intonation, employing both a high boundary tone and a raised global register. Here, the high boundary tone serves the dual purpose of signalling sentence mood (interrogativity) and marking structural boundaries, consistent with Pierrehumbert's (1980) theory of boundary marking and discourse interpretation. Acoustically, the high boundary tone manifests in GuanM as pitch compression for falling tones and a steeper trajectory for rising tones. This pattern is not unique to GuanM; C. Zhang (2018) observed a similar phenomenon in Tianjin Mandarin where the lexical tones

of the final syllable were not overridden and falling tones were deterred from falling fully while rising tones were facilitated higher, which led to the proposal of a floating boundary tone (H̰%).

In her doctoral research, C. Zhang (2018) conducted two production studies investigating the intonational yes-no question tune in Tianjin Mandarin. The first study examined monosyllabic words in isolation (Mono(ISO)), and the second investigated longer sentences where these same monosyllabic words served as sentence prominence (Mono(SEN)). She recorded six native Tianjin Mandarin speakers (three male, three female) aged 22-50, all of whom were born and raised in urban Tianjin. The materials consisted of 12 monosyllabic words using three syllables ([ma], [mi], [maʊ]) with four different lexical tones (L, H, LH, HL), recorded in both statements and questions.

Her research found that interrogative tunes in Tianjin Mandarin are characterised by three main features. First, yes-no questions had a consistent register lift, where both F0 maxima and minima were raised in questions compared to statements. Second, the lexical tone contours appeared to be preserved. In longer sentences, an H* pitch accent was associated with prominence in both tune types, followed by post-focus compression. Third, falling and low tones (L and HL) had narrower F0 ranges, while rising tones (H and LH) had wider F0 ranges in questions. Furthermore, questions were generally longer in duration at phrase-final positions. Based on these observations, C. Zhang proposed a global register lift combined with a H̰% for the question tune in Tianjin Mandarin. C. Zhang explicitly distinguished between "real" and "floating" boundary tones based on their phonetic realisation. She claimed that "a floating boundary tone should be different from a real boundary tone - a real boundary tone is an interpolation of an extra tone, while a floating boundary tone does not realise as an extra tone" (p. 129). Based on this definition, she argued that because the Tianjin interrogative tune did

not add a new tonal segment to the end of the phrase, but rather modified the range of the existing lexical tone, it could not be classified as a "real" boundary tone. Instead, she defined the floating boundary tone by its interaction with the neighbouring tone, suggesting that "a floating boundary tone should modify the neighbouring tone(s) without having its own phonetic realisation as a lexical floating tone does" (p. 129). Specifically, she proposed that "the 'benefit' of the floating boundary tone is that it both alters the post-lexical intonation and also keeps the lexical tone contour unchanged" (p. 191), acting to deter falling tones from falling and facilitate rising tones to rise higher. She further compared the interrogative tune with the Chanted Call (CC) tune, finding that the CC tune had "the L% boundary tone (which) is an interpolation of an extra L tone at the IP boundary and does not modify the neighbouring tones" (p. 191). Consequently, she concluded that "the L tone thus is not a 'floating' boundary tone but a real boundary tone" (p. 192).

In sum, C. Zhang's "floating" analysis rests on the premise that a "real" boundary tone must be an interpolated extra segment. However, this study adopts a different phonological interpretation, following the analysis established by Rialland and Aborobongui (2016) in their study of Embosi, which demonstrates that a boundary tone can be "real" and fixed even when it manifests through superimposition rather than interpolation. For instance, as shown in Figure 3.10, an L% boundary tone may be presented as a phonetically rising contour even when interacting with a high lexical tone, yet it remains a specific, local target. Crucially, the "floating" designation implies that the tone's position is not strictly measured or fixed at the edge. In contrast, the boundary tone in GuanM is fixed and aligned strictly to the edge of the IP, suggesting that the H% in GuanM is not a floating entity exerting general upward pressure, but a superimposed boundary tone, a local tone that is fully docked at the phrase edge and achieves its phonetic presence by modifying the trajectory of the final lexical tone.

If the boundary tone were a floating influence exerting a general upward pressure, the high-level tone T4 in this study should simply shift vertically upwards while maintaining its parallel, level trajectory. However, the GuanM data reveal a distinct positive reorientation of the second syllable of T4 trajectory: the contour shifts from slightly negative slope in statements to a positive inclination in questions, as shown in Table 3.5, and as shown in Figure 3.12, in which T4T4 with the question intonation has a significantly rising tendency. This demonstrates that the boundary tone does not merely lift the pitch; it actively changes the vector of the contour. The fact that the pitch bends upwards specifically at the phrase edge proves that the target is docked at the final coordinate of the intonational phrase. The trajectory is steering toward a specific high target, which necessitates a fixed association.

Following the superimposition relation between tone and intonation (Rialland & Aborobongui, 2016), we can explain the variation across the tone system without the need for additional categories or paralinguistic "strength" labels. In T4 (high-level), the last syllable lacks strong downward momentum, allowing the H% target to be fully realised. This creates a categorical slope reorientation from flat to rising. Meanwhile, in T1 (Falling), the lexical downward trajectory is robust, meaning the H% rise is physically constrained. The result is a mitigated fall, where the tone is prevented from falling fully, reaching a higher F0 ending point but still maintaining its falling shape.

Consequently, this study proposes that the high boundary tone (H%) in GuanM is an inherently rising phonological target aligned at the right edge of the IP. While its phonetic realisation may appear gradient, this is not due to a "weak" or "floating" status, but because the H% is physically constrained by the final lexical tone.

By adopting this model, we do not need to define extra entities or "floating" categories to account for phonetic variation. The apparent preservation of the falling tone is simply the phonetic byproduct of a fixed H% target interacting with a strong lexical filter. Unlike Cantonese, where boundary tones may appear as independent additive segments, the GuanM system is better understood as a categorical phonological target whose phonetic output is shaped by the lexical environment, allowing for a theoretically efficient model that accounts for both the fixed alignment and the gradient acoustic results.

3.5.7 Modelling Syntactically Unmarked Yes-No Questions

From the previous discussion, it is apparent that syntactically unmarked yes-no questions in GuanM are characterised by a global pitch raising and a phonologically High Boundary Tone (H%). Essentially, this H% influences the lexical tone contour not by altering its fundamental shape (i.e., by overriding it), but by being superimposed onto it. Consequently, the underlying tonal contours and tone dissimilation rules remain intact, even as the intonational contour modifies the overall pitch trajectory.

We then attempt to model yes-no questions by analysing the F0 difference, calculated by subtracting the statement F0 means from the question F0 means. Returning to the GAMM-fitted F0 contours (Figure 3.12), we observe that the F0 difference between questions and statements follows a nearly linear upward trend across all tones, which is substantial, ranging from a minimum of approximately 2 semitones to over 6 semitones at its peak. Furthermore, all three tones have an increasing difference over time, indicating that the distinction between questions and statements becomes more significant towards the end of utterances. Among them,

T1T1 shows the most dramatic difference, followed by T2T2, with T4T4 having the least difference.

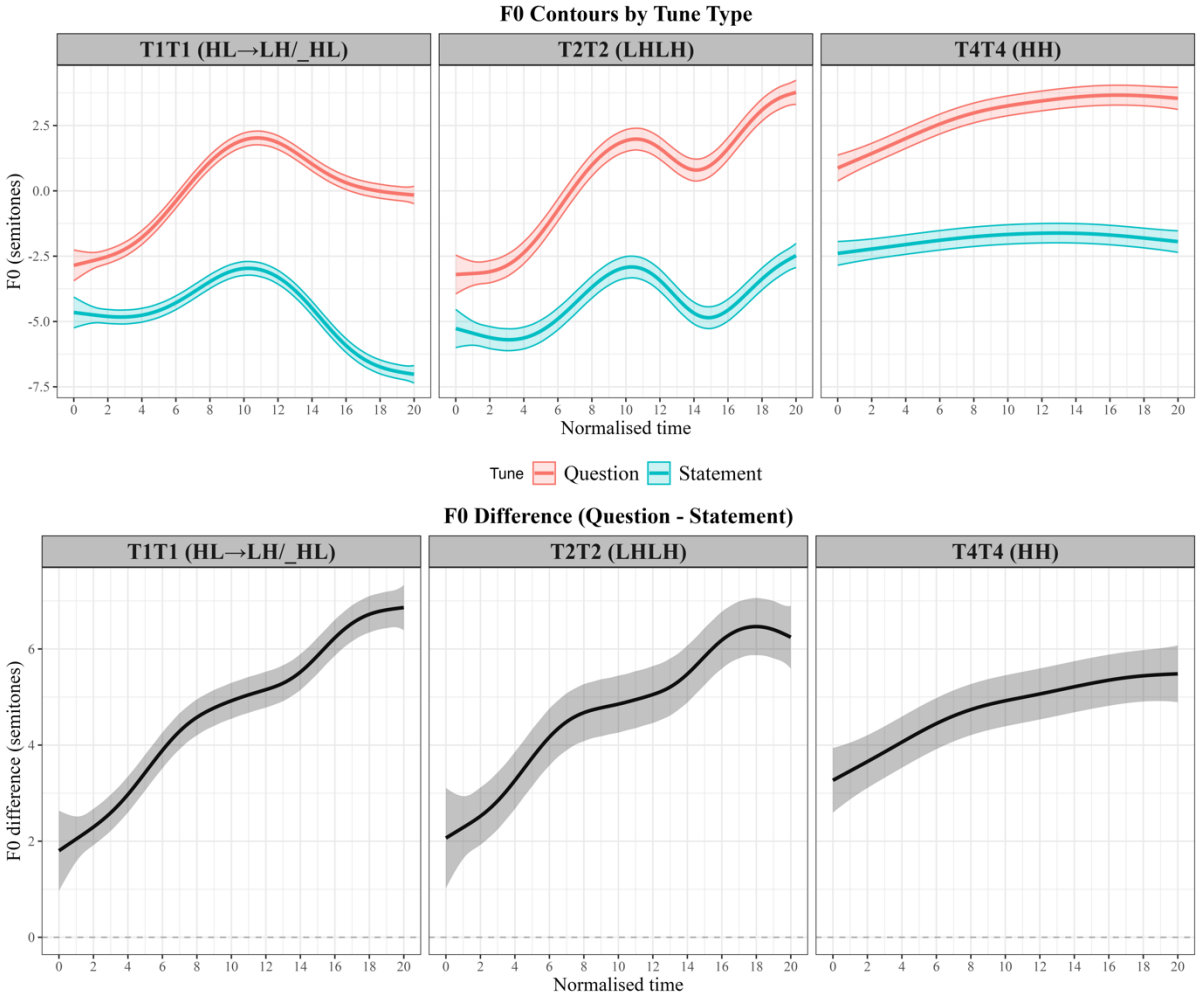


Figure 3.12 GAMM-fitted F0 contours for questions and statements (top panel) and the F0 differences between questions and statements (bottom panel)

The almost linear nature of the F0 difference between questions and statements provides strong empirical support for our proposed model of interrogative prosody. In subsequent sections, we will formalise the F0 patterns for syntactically unmarked yes-no questions in Mandarin languages as follows:

A consistent finding across all tones is the global register raising, which elevates the entire pitch contour of the intonational phrases:

$$R_Q = \delta(where\ \delta > 0)$$

in which register raising R_Q affects all aspects of pitch, increasing F0 mean, maximum, and minimum by δ.

In addition to the global raising, a high boundary tone is associated specifically with the final syllable of yes-no questions:

$$H_B(t) = \phi \frac{t}{d} (\phi > 0)$$

where ϕ is the scaling factor that determines the boundary tone strength, representing the total pitch increase contributed; t is the time within the final syllable (starting at 0 and ending at d); and d is the total duration of the final syllable. The ratio $\frac{t}{d}$ is the time-normalised progression within the final syllable, ranging from 0 to 1 across the final syllable, regardless of actual duration, meaning that at the syllable's start, $\frac{t}{d} = 0$ (no boundary tone effect) and at the end, $\frac{t}{d} = 1$ (maximum boundary tone effect).

This formulation captures the H% as a target attraction. It creates a gradient effect that increases towards the phrasal boundary, adding a constant positive slope component to the second syllable. Phonologically, this is critical: it accounts for how the H% preserves the identity of lexical tones while adjusting their phonetic realisation.

The formulation of the high boundary tone represents a rising contour rather than a high target because it fundamentally modifies the F0 change rate by adding a constant positive slope component to the second syllable. The phonetic effect manifests differently across tones, in which for level tones, it changes from zero slope to positive slope; for rising tones, it increases the existing positive slope; and for falling tones, it decreases the negative slope. This characterisation positions the boundary tone as a contour tone (defined by its movement pattern) rather than a register tone (defined by its relative height). Phonologically, this is important, since it better accounts for its interaction with different lexical tones, particularly how it preserves the identity of lexical tones while adjusting the phonetic realisation of the lexical tones on the surface form, so that the boundary tone creates a gradient effect that increases towards the phrasal boundary, and acts as an additive component to the pitch slope rather than just raising the register.

The F0 contour of an unmarked yes-no question is then represented as follows:

$$I_Q = R_Q + H_B(t) + T(t)$$

where R_Q represents the global register raising. $\phi \frac{t}{d}$ is the high boundary tone function. $T(t)$ is the lexical tone function.

Expanding on this, we get:

$$I_Q = R_Q + \phi \frac{t}{d} + T(t)$$

We assume only six basic lexical tone forms based on the register to simplify the analysis of the lexical tone function: High (Hr) or Low (Lr), and the contour shape: Rising (R), Falling (F), or Level (L). The contour shapes can be represented as linear functions as follows:

Level tone: $T_{X-L}\ (t) = a\ (constant)$

Rising tone: $T_{X-R}\ (t) = a + b \cdot t\ (b > 0)$

Falling tone: $T_{X-F}\ (t) = a - b \cdot t\ (b > 0)$

where a represents the starting pitch. b represents the slope (the rate of pitch change over time).

The table below illustrates the interaction between the high boundary tone and different final lexical tones in a syntactically unmarked yes-no question after applying the formula of I_Q above.

Table 3.17 Influence of the high boundary tone on final-syllable lexical tones in syntactically unmarked yes-no questions in GuanM

Final syllable tone	Phonetic surface	Description
$T_{Hr-L}\ (t)$ (High register, level contour)	Rising from high	The level tone shifts to a rising contour, beginning at a high pitch and rising further.
$T_{Hr-R}\ (t)$ (High register, rising contour)	Steep rising from high	Rise is exaggerated with a steeper slope ($b + \frac{\phi}{d}$), reaching an even higher peak
$T_{Hr-F}\ (t)$ (High register, falling contour)	Reduced falling from very high	Still falls but less steeply ($-b + \frac{\phi}{d}$); very high starting point
$T_{Lr-L}\ (t)$ (Low register, level contour)	Rising from mid	Level tone develops a rising contour, starting at a raised low position
$T_{Lr-R}\ (t)$ (Low register, rising contour)	Steep rising from mid	Enhanced rising with steeper slope ($b + \frac{\phi}{d}$); dramatic rise from the moderate start
$T_{Lr-F}\ (t)$ (Low register, falling contour)	Minimal falling or near-level	Significantly reduced falling; may appear almost level if $b \approx \frac{\phi}{d}$

Of particular note is the finding that high register tones with level or rising contours may show a more moderate boundary tone effect in terms of rising, which could be attributed to physiological constraints of the vocal tract, limiting the maximum pitch that can be produced. Such a constraint likely explains why T4 (high-level tone) in GuanM shows a relatively gradual rise in the differences of F0 between questions and statements, in comparison to T1T1 and T2T2.

3.6 Conclusion

This study employed a production experiment to investigate syntactically unmarked yes-no questions in disyllabic words as intonational phrases in GuanM. The findings suggest that in yes-no questions, lexical tone patterns (including rules for the lexical tone dissimilation) are applied within a local lexical domain, while intonation functions independently rather than overriding these tonal contours. More specifically, yes-no questions have an overall elevation in register compared to statements, featuring a phonological high boundary tone whose phonetic realisation varies depending on the lexical tone of the syllable it docks onto. Additionally, questions are produced with longer duration and higher intensity compared to statements.

The high boundary tone, located at the edge of the intonational phrase, is realised as a phonological rising contour from intonation. It interacts with the underlying lexical tones and modifies them in several ways. It enhances rising tones, reduces the steepness of falling tones, and converts level tones into rising tones. It maintains essential phonetic features of lexical tones, which secures their categorical identity, while it indicates interrogative mood in GuanM. Statements maintain a consistently lower register. The next step is to examine intonational patterns in longer utterances to test whether the patterns identified in this chapter are maintained.

Chapter 4 Production Study 2: Intonational Yes-No Questions in Guanzhong Mandarin in Focus

The previous chapter reported Production Study 1 on isolated productions. This chapter examines the production of longer utterances with focus in sentence-medial position, in both statement and question intonations. This chapter follows the structure of Chapter 3: it first reviews prosodic focus marking cross-linguistically, including Chinese languages, then addresses the research questions, and finally details the methodology used in this production study. The methods include stimuli preparation, the experimental procedure, and the statistical analysis of the acoustic features. It then discusses intonation in longer utterances in GuanM and the parallel encoding of focus and interrogative meanings before a concluding summary.

4.1 Prosodic Focus Cross-linguistically

Prosodically focused components are generally realised through pitch range expansion, prolongation, and increased intensity (e.g., Cooper et al., 1985; Heldner, 2003). For example, Féry and Kügler's (2008) study of German found that narrow focus raised pitch accents, making them more prominent. All-new sentences without explicit focus showed a downstepping pattern, where each subsequent high tone was progressively lower. In addition, narrowly focused constituents were significantly longer than their non-focused counterparts in all-new sentences. The focus-related lengthening was also found in English (Eady & Cooper, 1986). When a word was narrowly focused (e.g., due to a specific question prompting emphasis on that word), it exhibited higher F0 peaks and longer duration compared to neutral-focus sentences. In contrast, sentences with broad focus (e.g., focus on the entire verb phrase) showed an overall increase in duration, but unlike narrow focus, they did not exhibit a significant increase in F0 peaks.

The study of Chinese tones, tune and focus is complicated, as focus marking can be tone-dependent. By examining how different lexical tones were marked for focus, T. Wang et al. (2020) showed that in Mandarin Chinese, focus was primarily marked through pitch modifications, though the exact pattern depended on tone. For instance, focus raised the pitch register for T1, while T2 and T4 displayed prominent variations in pitch range and slope under focus. T3 distinctively lowered its pitch target when in focus, unlike the upward pitch trend observed in the other tones. Duration consistently increased for all focused tones, making it a strong cue for focus marking, while intensity played a minimal role.

Post-focus compression (PFC) refers to the reduction of pitch range and amplitude following a focused element. PFC has been reported in Indo-European, Altaic, and other languages, including Mandarin Chinese, but not in closely related Chinese languages, such as Cantonese and Taiwanese Hokkien (Y. Xu, 2011). Duan and Jia (2014) investigated PFC for focus realisation in declarative sentences across five Northern Mandarin dialects (Dalian, Harbin, Jinan, Tianjin, and Xi'an (GuanM)). Their results showed that F0 was the primary acoustic marker of focus, with all dialects exhibiting compressed pitch range in post-focus regions (i.e., PFC) while the pitch range remained unchanged in pre-focus areas. However, duration and intensity were not as significant as F0. Dialectal variations were observed in the realisation of on-focus F0: in Dalian, Jinan, and Tianjin, focus was marked by pitch range expansion across all on-focus syllables, whereas in Harbin and Xi'an, the second syllable of the focused word exhibited greater pitch expansion.

Chen (2010) questioned the assumption that PFC was a universal or primary characteristic of focus realisation in Mandarin Chinese. Instead, she observed that post-focus tones sometimes exhibited expanded rather than compressed F0 ranges and often displayed prolonged influence

from preceding tones, differing from typical coarticulation patterns. What consistently distinguished post-focus from on-focus tones was not pitch range but rather the reduced distinctiveness of tonal contours, especially following high tones. She proposed that post-focus effects should be understood as manifestations of weak tonal target implementation in prosodically non-prominent positions rather than as simple compression.

F. Liu and Xu (2005) examined the interaction of focus and interrogative meaning in Mandarin intonation. They found that both functions were encoded in parallel, allowing speakers to simultaneously express emphasis and question meaning without one obscuring the other. Focus was realised through pitch expansion on the focused word, post-focus compression, and unchanged pre-focus pitch, a pattern consistent across statements and questions. Additionally, focus interacted with question intonation, as questions containing focused elements exhibited an exponential rather than a linear pitch increase, with the largest difference occurring at the end of the sentence.

In Lhasa Tibetan, X. Zhang et al. (2012) found that focus was marked by rising F0, expanded pitch range, and lengthened duration, with these effects occurring in both statements and yes-no questions. Furthermore, PFC was observed, as the pitch of post-focus words was lowered and compressed, while pre-focus words remained largely unaffected. Yes-no questions were characterised by a higher F0 in unfocused words overall, but focused words showed no significant difference between the two sentence types.

Eady and Cooper (1986) stated that in English, focus had similar duration effects in yes-no questions and statements, with lengthening strictly applied to the focused word and not extending to others. They also confirmed the classical pattern of a rising terminal F0 contour

for questions and a falling contour for statements, regardless of the position of the focused words. The difference between questions and statements lay in post-focus prosody, depending on the focus position. When focus occurred at the beginning of a sentence, questions and statements differed dramatically: in statements, there was a sharp F0 drop after the focused word, resulting in a compressed and flattened contour for the rest of the utterance. In contrast, questions maintained raised F0 levels rather than lowering pitch throughout the post-focus region, although the pitch range remained compressed.

Such interaction between focus and sentence type does not occur in all languages. Z. Liu's (2018) study found little interaction between focus and interrogativity in Bai. The study concluded that focus and question intonation in Bai were independent. While focus was marked by duration lengthening of the focal element compared to post-focus elements, interrogative meaning was conveyed through lengthening duration, expanding the pitch span, and raising the F0 maximum and minimum of the sentence-medial constituents to distinguish from declaratives.

4.2 Research Questions

This study aims to explore the prosodic features associated with focus in GuanM, concentrating on the relationship between focus and sentence type (syntactically unmarked yes-no questions and statements). The first research question asks whether the global register and the high boundary tone are maintained in longer utterances in yes-no questions. The second research question seeks to determine which prosodic features, such as duration, intensity, or pitch, are most prominent and consistent in marking focus. Although earlier research has indicated that languages differ in their use of these acoustic features, it is still uncertain which cue is most reliable for marking focus in GuanM. A further question is how focus manifests differently in

questions compared to statements, particularly whether question marking and focus can be encoded in parallel through acoustic features, alongside tones.

4.3 Methods

4.3.1 Participants

Eleven native speakers of GuanM participated in this study (five females and six males; age range: 30 to 45 years; mean age: 36.63 years). All participants lived in Xi'an, China, and used GuanM as their main language in daily life. None of the participants reported any hearing or speech impairments. Informed written consent was obtained from all participants prior to the experiment.

4.3.2 Stimuli

The stimuli used in this study were the same as those in Production Study 1 (Table 3.3), consisting of 15 disyllabic words for each of the three tones: T1T1, T2T2, and T4T4. Participants were asked to produce these stimuli within longer utterances, with the target words placed in the focus position within either a question or statement tune. Each tune was embedded in a carrier sentence composed of six syllables: two for the pre-focus, two for the on-focus, and two for the post-focus phrases. The effects of preceding tones on the focus were manipulated using a disyllabic pre-target sequence: T1 (他/她 t^ha31 'he/she'), followed by one of four monosyllables with different tones - T1 (说 ʃuo31 'say'), T2 (读 tu24 'read'), T3 (写 siɛ52 'write'), and T4 (看 k^han55 'look at'). The post-focus phrase was the same in all conditions 那词 uo52 ts^hi24 'that word'. Each combination was repeated twice to minimise participant fatigue. Consequently, each participant recorded a total of 720 sentence configurations (4 pre-target syllables × 3 tones × 15 disyllabic words × 2 tunes × 2 repetitions).

4.3.3 Procedure

All audio recordings were made in a quiet room in the local community, using a Samson Meteor Mic connected to a Lenovo laptop with Audacity software, at a sampling rate of 44.1 kHz. The stimuli were presented to participants via Microsoft PowerPoint, with sentences centred on the slide. Participants were instructed to produce sentences according to the prompt type. In question prompts, participants were required to produce a confirmation-seeking question, while in statement prompts, they produced an affirmative response to a yes-no question.

For the question scenario, participants saw an affirmative response on the screen, such as 对，他/她说/读/写/看 [target word] 那词。tuei55, tʰa31 ʃuo31/tu24/siɛ52/kʰan55 [target word] uo52 tsʰi24 (‘Yes, he/she said/read/wrote/looked at [target word] that word’). Based on this affirmative answer, participants were required to create a corresponding yes-no question by asking, 他/她说/读/写/看 [target word] 那词？tʰa31 ʃuo31/tu24/siɛ52/kʰan55 [target word] uo52 tsʰi24 (‘Did he/she say/read/write/look at [target word] that word?’). This setup guided participants to form the appropriate interrogative question based on the provided affirmative response, as shown in (13)a. In the statement scenario, participants were presented with a yes-no question prompt, such as 他/她说/读/写/看 [target word] 那词？tʰa31 ʃuo31/tu24/siɛ52/kʰan55 [target word] uo52 tsʰi24 (‘Did he/she say/read/write/look at [target word] that word?’). Accompanying this prompt was the word “yes” in the answer, indicating that participants needed to produce a corresponding statement. Participants responded by saying, 对，他/她说/读/写/看 [target word] 那词。 tuei55, tʰa31 ʃuo31/tu24/siɛ52/kʰan55 [target word] uo52 tsʰi24 (‘Yes, he/she said/read/wrote/looked at [target word] that word’), as demonstrated in (13)b. Before the experiment, participants were given time to familiarise themselves with the stimuli and had 10 practice trials using words that were not included in the experimental stimuli. The entire session lasted approximately 45 minutes per participant, with breaks provided as needed.

(13)

a. Question scenario

Participant response:

__?

Prompt on screen:

对，他/她说/读/写/看 [target word] 那词。

tuei55, t^ha31 ʃuo31/tu24/siɛ52/k^han55 [target word] uo52 ts^{h}i24。

'Yes, he/she said/read/wrote/looked at [target word] that word.'

b. Statement scenario

Prompt on screen:

他/她说/读/写/看 [target word] 那词？

t^ha31 ʃuo31/tu24/siɛ52/k^han55 [target word] uo52 ts^{h}i24？

'Did he/she say/read/write/look at [target word] that word?'

Participant response:

对，__。

tuei55, __。

'Yes, __.'

4.3.4 Annotation and Alignment

Hesitations and misread sentences were excluded from the data analysis, resulting in approximately 5.27% data loss. In total, 7,503 sentences were used for further analysis, which included 3,697 statements and 3,806 questions. The alignment criteria were consistent with those used in Production Study 1, to ensure that both statements and questions had the same length of six syllables. The prompt 对 [tuei55] 'yes' was recorded but excluded in the annotated files. The audio files were annotated with three tiers: a syllable tier, a phrasal tier (divided into pre-focus, on-focus, and post-focus phrases), and a sentence-level tier.

4.3.5 Measurement

Acoustic measurements were taken at two levels (the phrasal tier and the on-focus disyllabic word) and included duration, mean intensity, mean F0, maximum F0, minimum F0, and F0 range. The timing of the maximum and minimum F0 was also measured.

4.3.6 Statistical Analyses

F0 was extracted from six-syllable utterances and segmented into 10 equal intervals per syllable, yielding a total of 60 time points per token. Pitch was converted to semitones relative to each speaker's mean F0, using the same normalisation method employed in Production Study 1. Generalised additive mixed models (GAMMs) were fitted using the mgcv R package (Wood, 2011) to model nonlinear pitch contours and account for variability across speakers and items. Semitone-normalised F0 was modelled as smooth functions of normalised time, with an interaction term between Tone, Tune, and Pre-target. Random smooths were included for Speaker and Item to accommodate inter-speaker and item-level differences.

For the statistical analysis of acoustic measurements, linear mixed-effects models were fitted using the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2024) to examine the interaction between Tune, Tone, Pre-target, and Phrase or Syllable across multiple acoustic variables, including Duration, Mean Intensity, Mean F0, Maximum F0, Minimum F0, and F0 Range. Two separate analyses were conducted: one at the phrasal level (pre-focus, on-focus, and post-focus phrases) and another focusing specifically on the syllables within the on-focus position (Syllable 1 and Syllable 2). For each acoustic parameter, the models included a four-way interaction between Tune, Tone, Pre-target, and either Phrase or Syllable. Random effects included Speaker, Item, and Repetition. The models are shown in (14):

(14) a. At the phrase-level

lmer(Acoustic_parameter ~ Tune * Tone * Phrase * Pre-target + (1|Speaker) + (1|Repetition) + (1|Item))

b. At the syllable-level of the on-focus phrases

lmer(Acoustic_parameter ~ Tune * Tone * Syllable * Pre-target + (1|Speaker) + (1| Repetition) + (1|Item))

Following the main analyses, two types of post-hoc comparisons were conducted using the emmeans R package (Lenth, 2024) with Tukey's adjustment for multiple comparisons. The first set of comparisons examined the effects of Tune (Question vs. Statement) while accounting for all other factors, including Pre-target. The second set of post-hoc tests averaged across Pre-target conditions, focusing solely on the interactions between Tune, Tone, and position factors (either Phrase or Syllable).

4.4 Results

4.4.1 Results of GAMMs

Figure 4.1 to Figure 4.3 provide examples of waveforms and spectrograms for one stimulus from each of the three tone categories, showing statement productions (left) and question productions (right) from the same female speaker. In these figures, the pre-targets are all T1 [ʃuo31].

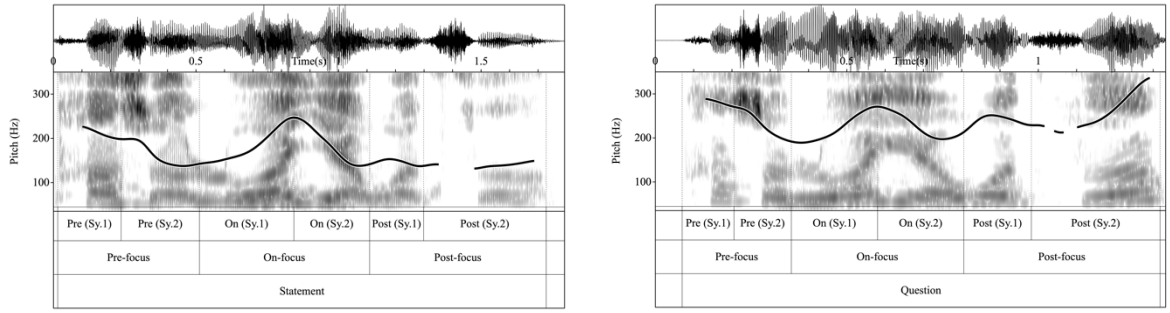


Figure 4.1 The statement tune (left panel) and question tune (right panel) of T1T1

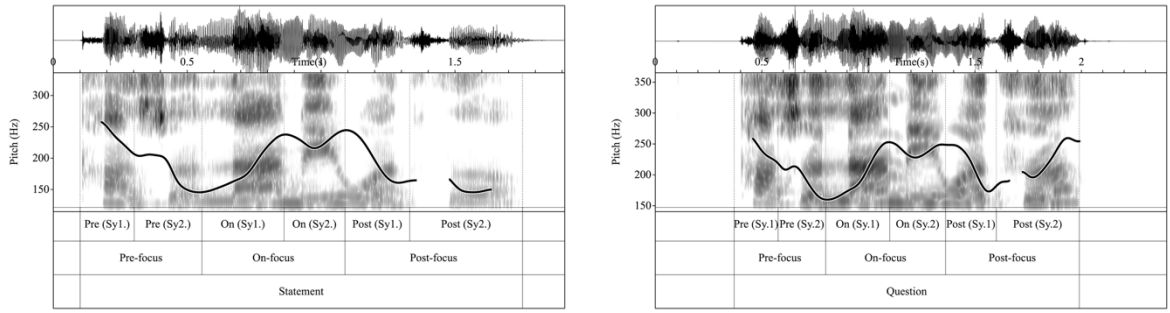


Figure 4.2 The statement tune (left panel) and question tune (right panel) of T2T2

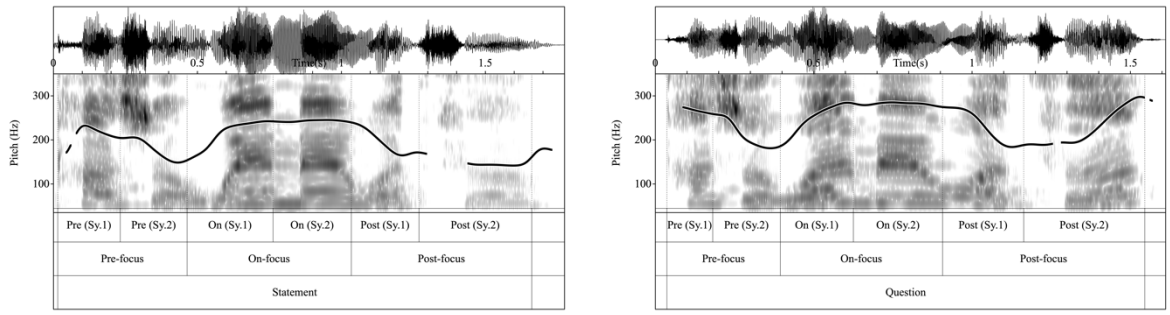


Figure 4.3 The statement tune (left panel) and question tune (right panel) of T4T4

The following section presents GAMM-fitted F0 contour plots at the sentence level to illustrate global trends, followed by a more detailed analysis across three prosodic domains: pre-focus, on-focus, and post-focus. Figure 4.4 displays the time-normalised F0 contours of full sentences

across three tonal combinations (T1T1, T2T2, T4T4), each featuring different pre-target tones. A clear overall pattern is observed: the F0 contours of questions consistently remain above those of statements throughout the utterance. In other words, question tunes maintain a higher pitch register than their statement counterparts across all three phrases.

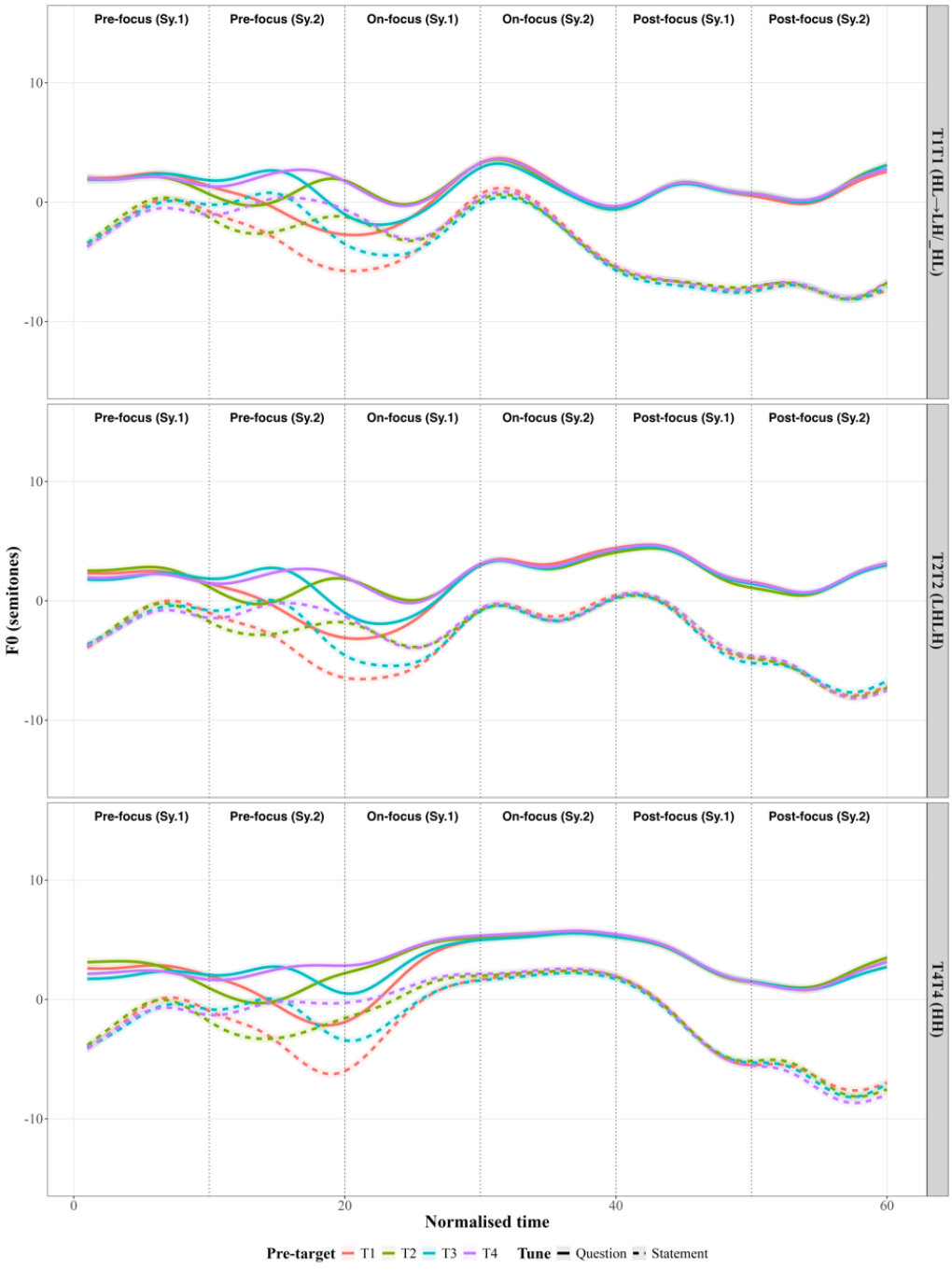


Figure 4.4 Time-normalised smoothed F0 contours of sentences with pre-target tones across three tones, fitted using GAMMs with 95% confidence intervals

Figure 4.5 below shows the F0 differences between question and statement contours with pre-target tones included. Across all tones and phrases, the F0 differences are consistently positive, implying that question intonation is realised with a higher pitch than statements globally. In the pre-focus phase, the distinction between the two intonations remains minor and consistent, generally 2-3 semitones. A slight rise is observed in the on-focus phrase, yet the pitch difference remains relatively modest, indicating only a subtle uplift of question contours compared to statements. The most significant distinction arises in the post-focus phrase, where the F0 of the question intonation rises steeply, often reaching its peak towards the end of the utterance. This rise can reach 8-10 semitones or more, depending on the preceding tones in the focus phrases.

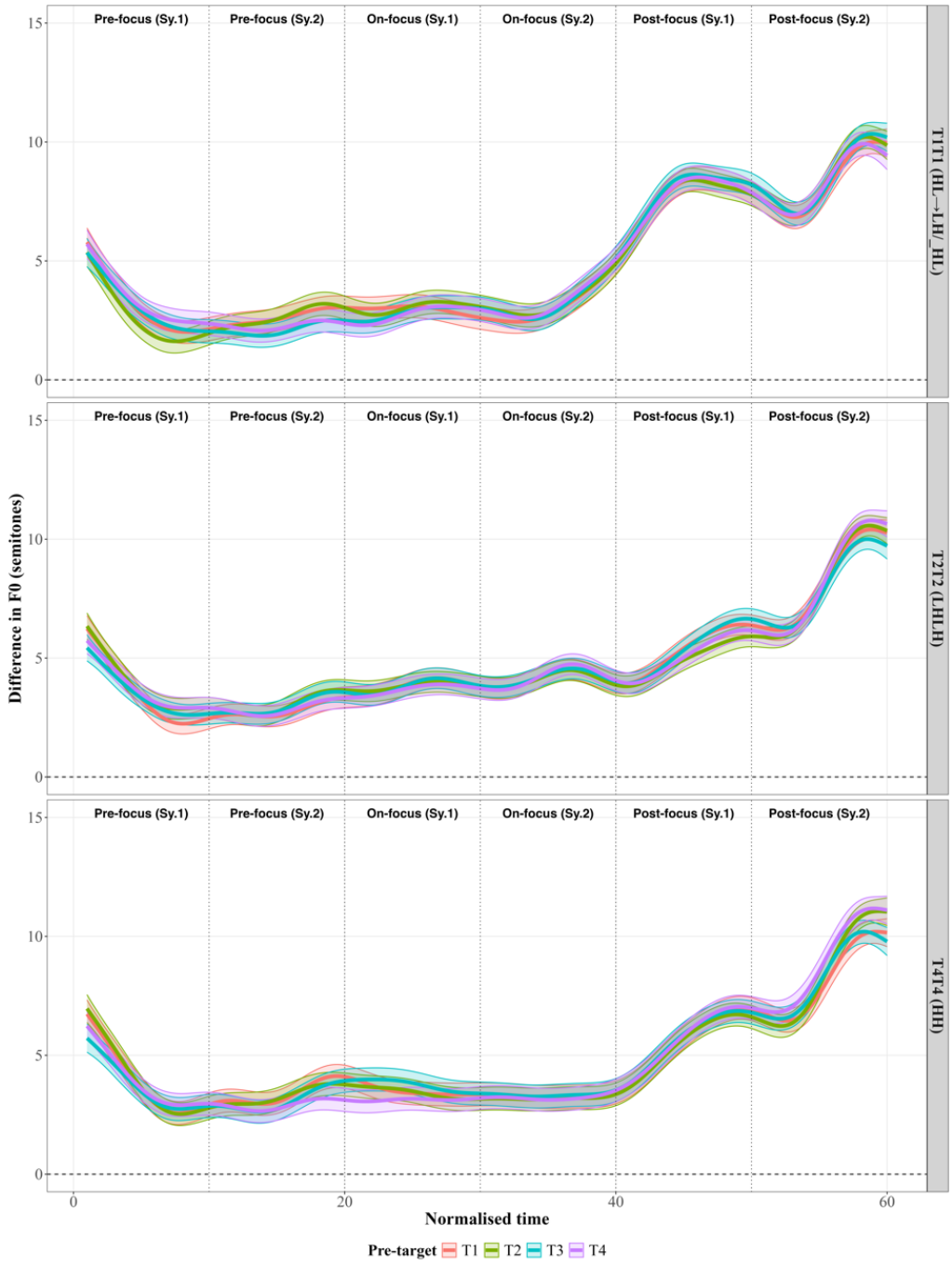


Figure 4.5 Time-normalised smoothed F0 differences of sentences across three tones with pre-target tones, fitted using GAMMs with 95% confidence intervals

Figure 4.6 below presents the F0 contours of questions and statements (top panel), as well as their respective F0 differences (bottom panel) with four pre-target tones in the pre-focus phrase. Although the contours for both tunes maintain a largely parallel relationship across all three

tones with F0 contours of questions being posited above the corresponding statement F0 contours, each pre-target tone displays its distinctive pitch shape on the second syllable. T1 and T3 exhibit falling contours but differ in pitch registers (low versus high), T2 demonstrates a rising pattern, and T4 presents as a high-level tone.

Despite these differences, two consistent patterns are observed. First, questions consistently maintain a higher register than statements throughout the pre-focus phrase. This raised pitch begins from the very onset of interrogative utterances, suggesting speakers plan the question intonation from the start. Second, statements show contours that begin slightly lower, rise to an early peak, and then gradually descend.

The F0 difference between questions and statements (bottom panel) initially shows a slight dip followed by stabilisation, reflecting a consistent register-level distinction between the two tunes that is established early and maintained throughout the pre-focus phrase, regardless of the specific lexical tones involved.

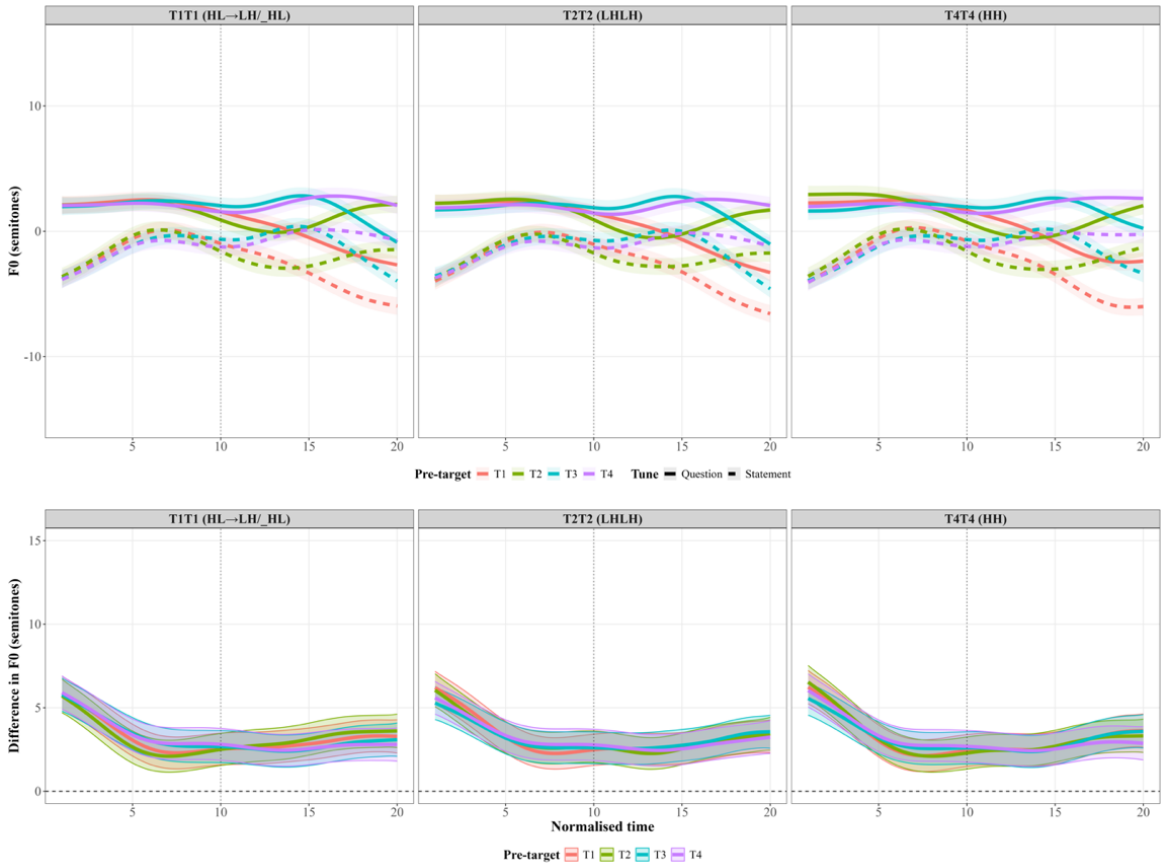


Figure 4.6 Time-normalised smoothed F0 contours (top panel) and F0 differences (bottom panel) of the pre-focus phrase with pre-target tones across three tones, fitted using GAMMs with 95% confidence intervals

Figure 4.7 displays the GAMM-fitted F0 contours for the on-focus phrase. As observed in the pre-focus phrase, question tunes consistently exhibit higher pitch values than statements across all tones throughout the entire phrase. In the top panel, the contours for questions are systematically positioned above those for statements, regardless of tone or pre-target tonal context. Furthermore, the raised pitch register in questions appears stable and is not significantly influenced by the lexical tone preceding the focused phrase. Additionally, the underlying lexical tonal contours remain intact within both sentence types. For instance, T1T1 displays a rising-falling pattern, consistent with the application of tone dissimilation; T2T2

shows a steady rising shape; and T4T4 maintains a relatively flat, high-level contour. The lower panel confirms that the F0 difference between question and statement tunes remains steady across time and tones, in contrast to the findings from Production Study 1, where disyllabic words as whole intonational phrases exhibited increasing F0 differences towards the end (as shown in Figure 3.12). The different F0 patterns suggest that pitch differences are largely attributable to the global register height in the focus phrase.

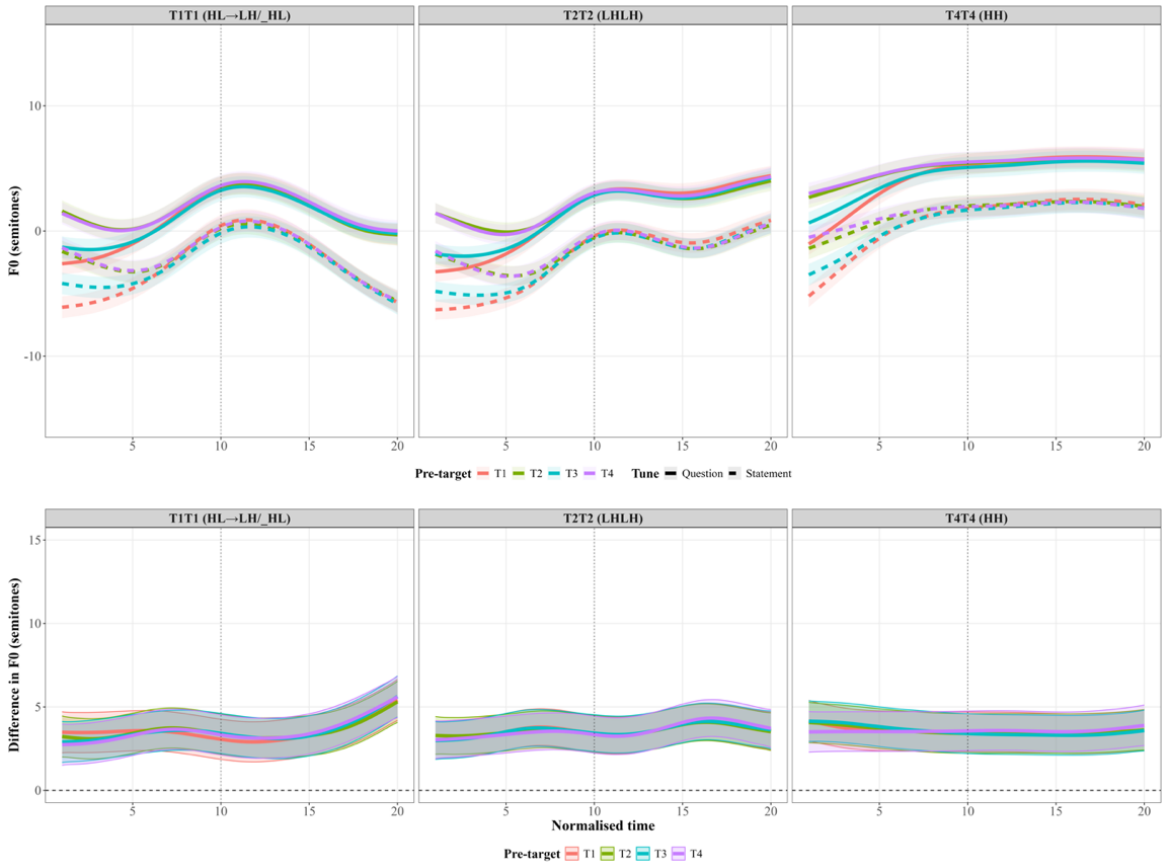


Figure 4.7 Time-normalised smoothed F0 contours (top panel) and F0 differences (bottom panel) of the on-focus phrase with pre-target tones across three tones, fitted using GAMMs with 95% confidence intervals

Figure 4.8 presents the post-focus phrases, each containing the same disyllabic word bearing an HL tone followed by an LH tone. Across all tones, the F0 of questions remains consistently higher than statements. In this phrase, question contours show a slightly falling pitch in the first syllable, followed by a clear rising trend in the second syllable. Thus, the pitch trajectory of questions across the whole phrase is either slightly falling or even slightly rising, particularly following T1T1 in the on-focus phrase, which reflects the carryover effect of T1 as a low falling tone. In contrast, as for the F0 contours of statements, while the second syllable may rise slightly, the overall F0 contour shows a steady downward trend, characteristic of declination. In the lower panel, the F0 differences between question and statement tunes are again uniformly positive. The difference is quite consistent in the first syllable (approximately 5-6 semitones) but rises significantly as the phrase concludes, which implies that there is a high boundary tone in questions, resulting in a more prominent pitch increase on the last syllable, with the most significant pitch difference noted at the end of the sentence, along with a global high register.

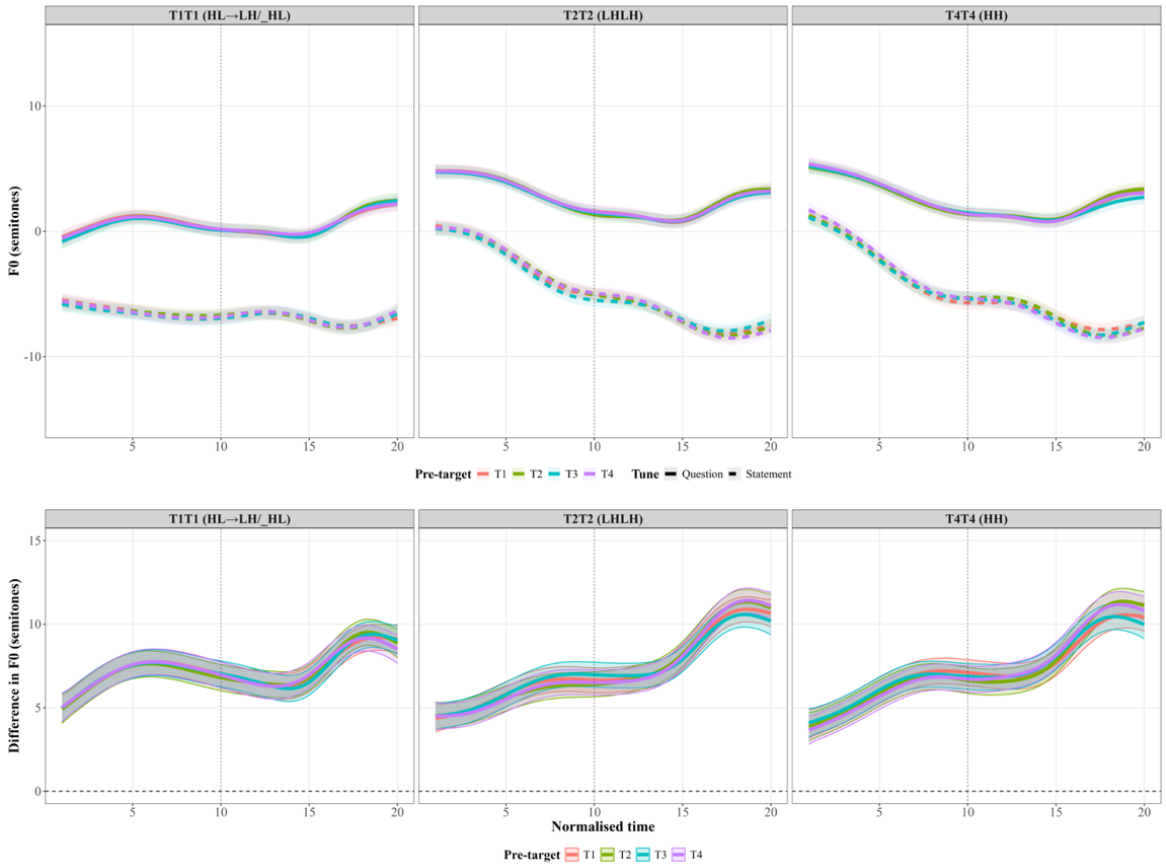


Figure 4.8 Time-normalised smoothed F0 contours (top panel) and F0 differences (bottom panel) of the post-focus phrase with pre-target tones across three tones, fitted using GAMM with 95% confidence intervals

4.4.2 F0 Change Rate

The F0 change rate is computed as the first derivative of the GAMM-fitted model's F0 trajectories, reflecting how rapidly the F0 is rising or falling at each point in time and is expressed in semitones per unit of normalised time. Figure 4.9 shows the F0 change rates, averaged across six syllable positions, in which red cells indicate rising pitch contours (positive slope), blue cells represent falling contours (negative slope), and white cells correspond to relatively flat contours.

In the pre-focus region for questions, the first syllables of the whole utterances frequently have negatives slopes, showing a downward pitch movement that is indicative of a localised F0 dip early in the utterance. This may reflect anticipatory intonational structuring, where pitch is lowered in preparation for a subsequent pitch rise associated with the interrogative contour. Following the initial dip, the pitch movement trajectory varies based on the pre-target tone: T1 and T3 exhibit falling contours, T2 is rising, and T4 is level. In statements, the first syllables show a mild upward pitch movement, aligning with the early rise often observed at the beginning of declarative utterances, marking the onset of the intonational phrase. The second syllables preserve the lexical contour shape of the pre-target tones, as is observed in questions.

In the on-focus region, the F0 change rate of both statements and questions strongly reflect the lexical tone contours, regardless of pre-target tones. For example, T1T1, a rising-falling contour, shows a positive F0 change in the first focused syllable, followed by a negative change in the second in both tunes. Furthermore, when the pre-target tone is T1, the T1T1 sequence in the on-focus phrase does not change the contour shape, meaning that the lexical tonal alternation rule of T1T1 is only applied to the lexical level and does not cross the phrasal boundary.

The post-focus phrases of T2T2 and T4T4 have a similar pattern in both questions and statements: the slopes of the first syllables in both tunes are negative, indicating that from the high tone in the on-focus phrase to the high tone in the post-focus region, there is a falling tendency. In the second syllable, the slopes of questions are positive, indicating a rising tendency linked to the high boundary tone, while in statements, the slopes remain negative, suggesting a continuation of the overall downward trend of the statement tune. This pattern appears to be reversed in T1T1. In questions, both syllables show positive slopes, suggesting

that the falling tone (T1) in the on-focus phrase sets a low pitch baseline at the beginning of the post-focus phrase, after which the pitch rises, likely influenced by the high boundary tone. In statements, the falling tendency of the preceding T1 tone also sets the pitch low. With the additional declination brought about by the statement intonation pattern, the pitch continues to fall. However, since the pitch level is already relatively low, the second syllable shows a slightly positive slope, indicating a modest rising movement, though considerably weaker than the question tune.

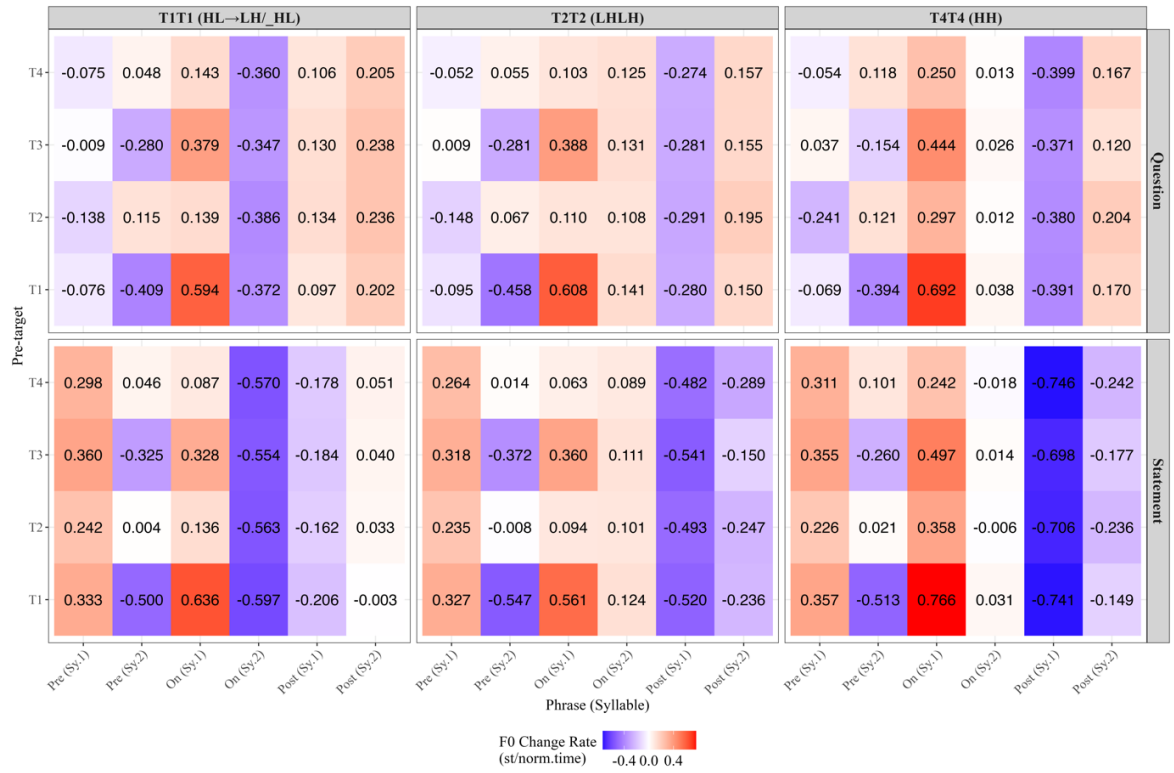


Figure 4.9 Average F0 change rates (in semitones per normalised time) across phrase

We then move to the F0 change rate calculated by entire utterances, separated by tone, tune and pre-target tone, as shown in Figure 4.10 below. Statements consistently have negative slopes across all tonal combinations, as evidenced by the uniformly blue-shaded cells in the lower portion of the figure. These negative values, typically ranging from -0.048 to -0.066

semitones per normalised time, indicate a decreasing F0 pattern that aligns perfectly with the characteristic pitch decline observed in declarative utterances. Conversely, questions uniformly exhibit positive average F0 slopes, represented by the red-highlighted cells in the upper portion of the figure. These positive values, varying from approximately 0.006 to 0.021, demonstrate a clear upward trend in F0 during interrogative utterances, which persists regardless of the lexical tone of the emphasised syllable or the pre-target tone, suggesting that the overall pitch rise in questions constitutes the intonational feature that transcends local tonal effects.

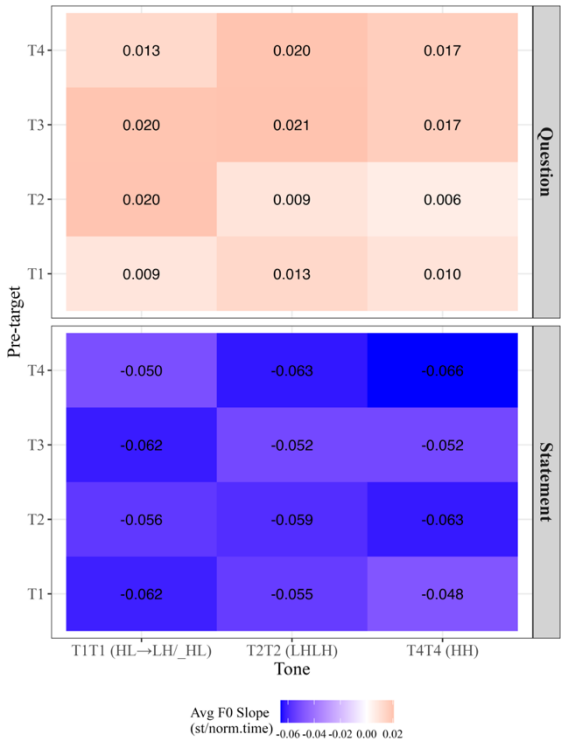


Figure 4.10 F0 change rates (in semitones per normalised time) at the utterance-level by tone, tune and pre-target tone

The following sections present the results for duration, mean F0 intensity, mean F0, maximum F0, minimum F0, and F0 range at both the phrasal level and the disyllabic level of the on-focus phrases. Finally, the maximum F0 and minimum F0 alignment is presented.

4.4.3 Duration

4.4.3.1 Phrase-Level

Significant main effects were found for Tune ($F(1, 21{,}482.08) = 90.35$, $p < .001$), Tone ($F(2, 41.69) = 3.41$, $p = .04$), Phrase ($F(2, 21{,}471.60) = 6{,}764.00$, $p < .001$), and Pre-target tone ($F(3, 21{,}478.87) = 23.52$, $p < .001$). Significant two-way interactions were observed between Tune and Phrase ($F(2, 21{,}471.60) = 268.81$, $p < .001$), Tone and Phrase ($F(4, 21{,}471.60) = 19.37$, $p < .001$), and Phrase and Pre-target tone ($F(6, 21{,}471.60) = 12.80$, $p < .001$). No significant effects were found for interactions between Tune and Tone, Tune and Pre-target tone, Tone and Pre-target tone, or for any higher-order interactions (ps > .05) (See Table 4.1).

Table 4.1 Effects of Tune, Tone, Phrase, and Pre-target on Duration[11]

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	501,740.40	501,740.40	1	21,482.08	90.35	< .001***
Tone	37,872.65	18,936.33	2	41.69	3.41	.0425*
Phrase	75,123,221.40	37,561,610.70	2	21,471.60	6,764.00	< .001***
Pre-target	391,863.67	130,621.22	3	21,478.87	23.52	< .001***
Tune:Tone	4,691.49	2,345.74	2	21,481.13	0.42	.6555
Tune:Phrase	2,985,463.07	1,492,731.54	2	21,471.60	268.81	< .001***
Tone:Phrase	430,361.86	107,590.47	4	21,471.60	19.37	< .001***
Tune:Pre-target	9,561.25	3,187.08	3	21,480.59	0.57	.6321
Tone:Pre-target	31,379.71	5,229.95	6	21,478.87	0.94	.4634
Phrase:Pre-target	426,571.64	71,095.27	6	21,471.60	12.80	< .001***
Tune:Tone:Phrase	21,056.73	5,264.18	4	21,471.60	0.95	.4349
Tune:Tone:Pre-target	26,653.06	4,442.18	6	21,480.53	0.80	.5698
Tune:Phrase:Pre-target	66,235.20	11,039.20	6	21,471.60	1.99	.0637
Tone:Phrase:Pre-target	57,567.32	4,797.28	12	21,471.60	0.86	.5838
Tune:Tone:Phrase:Pre-target	43,174.26	3,597.86	12	21,471.60	0.65	.8024

Post-hoc comparisons showed that questions consistently had shorter durations in both the pre-focus and on-focus positions compared to statements, regardless of tone and pre-target tone,

[11] In the table headings and hereinafter, SS = sum of squares, MS = mean square, df (Num) = numerator degrees of freedom, df (Den) = denominator degrees of freedom, *F* = *F*-statistic, *p* = *p*-value.

with all differences statistically significant (*p*s < .05). Questions consistently showed longer durations in the post-focus position than statements (*p*s < .001). Figure 4.11 presents the effect of tunes on duration with longer duration in questions by the estimated marginal means of duration across different phrases (pre-focus, on-focus, post-focus) for each tone, separated by pre-target tones.

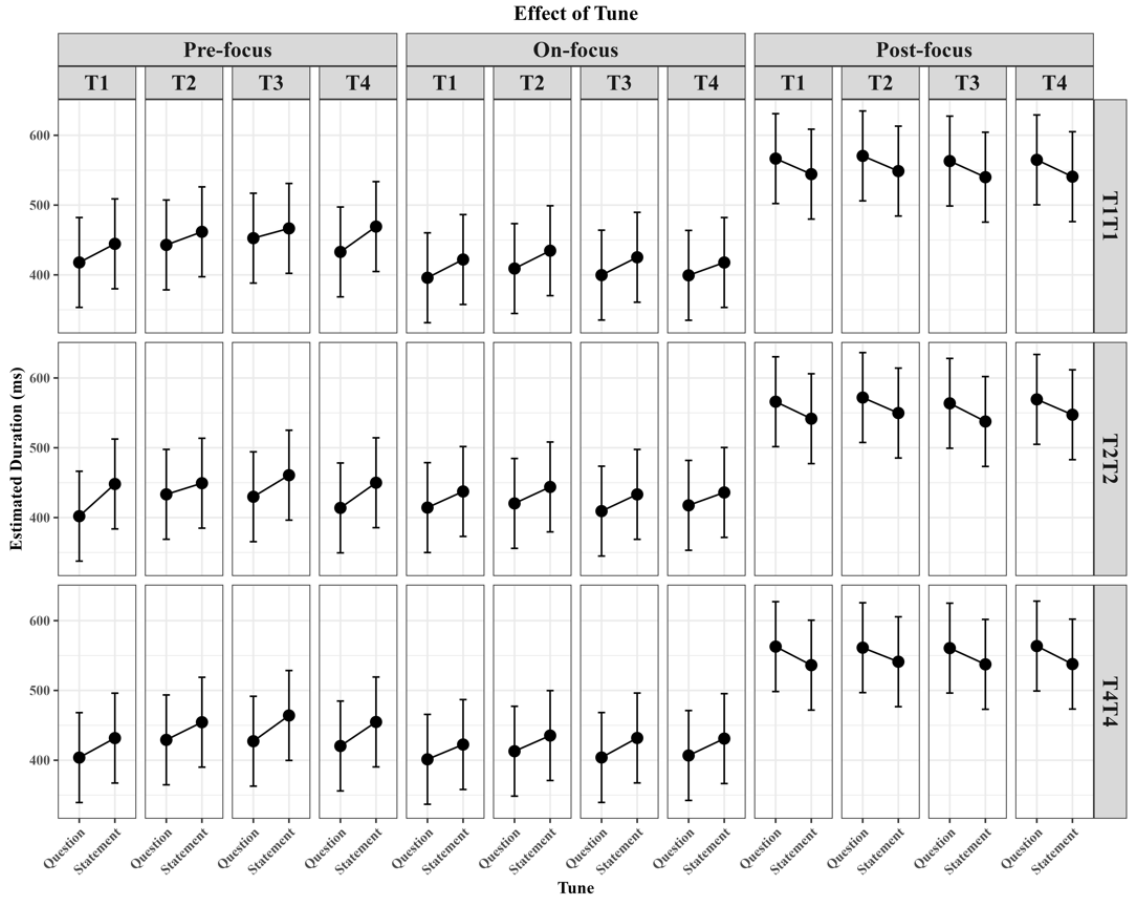


Figure 4.11 Effect of Tune on Duration across pre-target tones: Estimated marginal means with 95% confidence intervals

Figure 4.12 shows the effect of phrases on duration by the estimated marginal means of duration across different tunes for each tone, separated by pre-target tones, with error bars representing 95% confidence intervals. Irrespective of tones and pre-target tones, duration was

consistently observed across both sentence types for most tonal and pre-target tone combinations in the following order: post-focus > pre-focus > on-focus, with the on-focus phrase typically being the shortest. Duration in the post-focus position was significantly longer than in both the pre-focus and on-focus positions (all *p*s < .001), regardless of tone, tune, or pre-target tone.

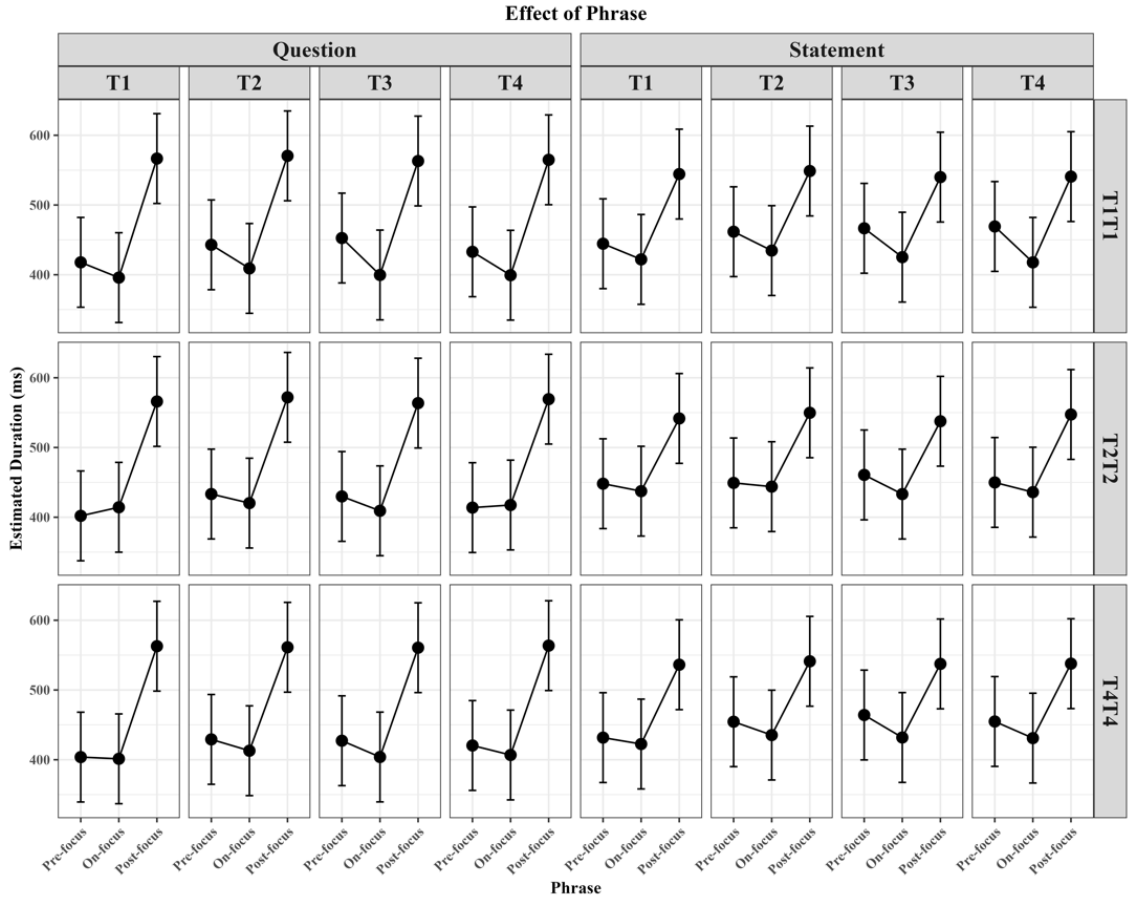


Figure 4.12 Effect of Phrase on Duration across pre-target tones: Estimated marginal means with 95% confidence intervals

The influence of pre-target tone was primarily found in both pre-focus and on-focus positions, suggesting both anticipatory and carryover effects. To be specific, in the pre-focus position, the most considerable durational contrasts between questions and statements occurred when

the upcoming tone was T2T2 in the on-focus position, particularly when preceded by T1 pre-target tone (a low falling tone; β = -46.10, t(21475) = -7.63, p < .0001). Substantial effects were also observed in conditions involving T4 pre-target tone (a high-level tone), such as followed by T1T1 (β = -36.30, t(21475) = -5.92, p < .0001) and T2T2 (β = -36.10, t(21478) = -5.92, p < .0001). Pre-target tone has a possible anticipatory effect on duration, where upcoming tones influence the duration of preceding phrases.

Pre-target tone also influences the duration of the on-focus position, demonstrating carryover effects from preceding tones. Although on-focus durations were consistently shorter than adjacent regions across different tones and tunes, the extent of the duration compression varied depending on pre-target tone. For example, when the pre-target tone was T1, on-focus question-statement differences ranged from β = -21.20 to -26.10 across tones. The T3 pre-target tone (a high falling tone) led to powerful effects in the on-focus region, especially when followed by T4T4 (β = -28.00, t(21474) = -4.59, p < .0001) and T1T1 (β = -25.60, t(21475) = -4.18, p < .0001). See Table A.1 in Appendix A for the mean duration.

4.4.3.2 Syllable-Level (On-Focus)

The table below (Table 4.2) presents the results of the main effects at the syllable level in the on-focus phrases. Tune exhibited a highly significant main effect on duration (F(1, 14,343.08) = 375.56, p < .001), as well as Syllable (F(1, 14,341.95) = 2,128.49, p < .001), and Pre-target tone (F(3, 14,342.65) = 15.03, p < .001). No significant main effect was found for Tone (F(2, 43.98) = 2.36, p = .1065). Among the interaction terms, significant effects were observed among the interaction terms: Tune × Syllable (F(1, 14,341.95) = 32.48, p < .001), Tone × Syllable (F(2, 14,341.95) = 168.28, p < .001), Tone × Pre-target (F(6, 14,342.65) = 2.35, p = .0286), Syllable × Pre-target (F(3, 14,341.95) = 7.50, p < .001), and Tone × Syllable × Pre-

target ($F(6, 14{,}341.95) = 2.11$, $p = .0490$). No other interactions reached statistical significance (all ps > .05).

Table 4.2 Effects of Tune, Tone, Syllable, and Pre-target on Duration

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	477,136.79	477,136.79	1	14,343.08	375.56	< .001***
Tone	5,990.46	2,995.23	2	43.98	2.36	.1065
Syllable	2,704,175.27	2,704,175.27	1	14,341.95	2,128.49	< .001***
Pre-target	57,298.89	19,099.63	3	14,342.65	15.03	< .001***
Tune:Tone	821.24	410.62	2	14,342.89	0.32	.7238
Tune:Syllable	41,268.02	41,268.02	1	14,341.95	32.48	< .001***
Tone:Syllable	427,587.87	213,793.93	2	14,341.95	168.28	< .001***
Tune:Pre-target	3,831.01	1,277.00	3	14,342.82	1.01	.3893
Tone:Pre-target	17,911.95	2,985.32	6	14,342.65	2.35	.0286*
Syllable:Pre-target	28,600.70	9,533.57	3	14,341.95	7.50	< .001***
Tune:Tone:Syllable	6,208.25	3,104.12	2	14,341.95	2.44	.0869
Tune:Tone:Pre-target	3,762.25	627.04	6	14,342.82	0.49	.8137
Tune:Syllable:Pre-target	3,023.29	1,007.76	3	14,341.95	0.79	.4975
Tone:Syllable:Pre-target	16,068.65	2,678.11	6	14,341.95	2.11	.0490*
Tune:Tone:Syllable:Pre-target	7,451.20	1,241.87	6	14,341.95	0.98	.4385

The effect of Tune was consistent across three tones, with questions yielding significantly shorter syllable durations than statements, with the difference of duration between questions and statements being particularly larger for Syllable 1. For instance, when Tone = T1T1 and pre-target tone = T2, the difference in duration between question and statement was -20.85 ms for Syllable 1 ($t(14{,}343) = -7.05$, $p < .0001$), compared to a non-significant -3.85 ms for Syllable 2 ($t(14{,}343) = -1.30$, $p = .19$). A similar pattern was observed for Tone = T2T2, pre-target tone = T3, where Syllable 1 showed a significant difference of -17.94 ms ($t(14{,}342) = -6.21$, $p < .0001$), while Syllable 2 exhibited a marginal effect of -5.30 ms ($t(14{,}342) = -1.84$, $p = .07$).

Syllable 1 was consistently and significantly longer than Syllable 2 across all tone and pre-target tone combinations. The extent of this difference was modulated by both tone identity and tune type. For example, under Tone = T1T1, pre-target tone = T3, and Tune = Statement, the contrast reached 41.50 ms (t(14,342) = 13.85, $p < .0001$), whereas under T4T4 with the same pre-target tone and Tune, the difference was markedly smaller at 13.80 ms (t(14,342) = 4.64, $p < .0001$). The detailed mean duration across tone, syllable and pre-target tone is listed in Table A.2 in Appendix A.

4.4.4 Mean Intensity

4.4.4.1 Phrase-Level

A strong main effect of Tune (F(1, 21,474.24) = 9,910.26, $p < .001$) was found that questions were produced with significantly higher intensity than statements in each phrase position across all tones and pre-target tones (all ps < .0001). See Table 4.3 below, and Figure 4.13 which shows the effect of Tune on mean intensity, with such observations as the estimated marginal means across different phrases for each tone, separated by pre-target tones.

Table 4.3 Effects of Tune, Tone, Phrase, and Pre-target on Mean Intensity

Factor	SS	MS	df (Num)	df (Den)	F	p
Tune	21,490.88	21,490.88	1	21,474.24	9,910.26	< .001***
Tone	111.82	55.91	2	42.00	25.78	< .001***
Phrase	73,266.14	36,633.07	2	21,471.98	16,892.90	< .001***
Pre-target	248.22	82.74	3	21,473.30	38.15	< .001***
Tune:Tone	1,233.41	616.71	2	21,473.75	284.39	< .001***
Tune:Phrase	13,900.26	6,950.13	2	21,471.98	3,204.97	< .001***
Tone:Phrase	2,269.77	567.44	4	21,471.98	261.67	< .001***
Tune:Pre-target	85.75	28.58	3	21,473.63	13.18	< .001***
Tone:Pre-target	18.19	3.03	6	21,473.30	1.40	.2111
Phrase:Pre-target	846.25	141.04	6	21,471.98	65.04	< .001***
Tune:Tone:Phrase	1,283.12	320.78	4	21,471.98	147.92	< .001***
Tune:Tone:Pre-target	23.03	3.84	6	21,473.63	1.77	.1009
Tune:Phrase:Pre-target	122.67	20.44	6	21,471.98	9.43	< .001***

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tone:Phrase:Pre-target	21.35	1.78	12	21,471.98	0.82	.6296
Tune:Tone:Phrase:Pre-target	14.66	1.22	12	21,471.98	0.56	.8730

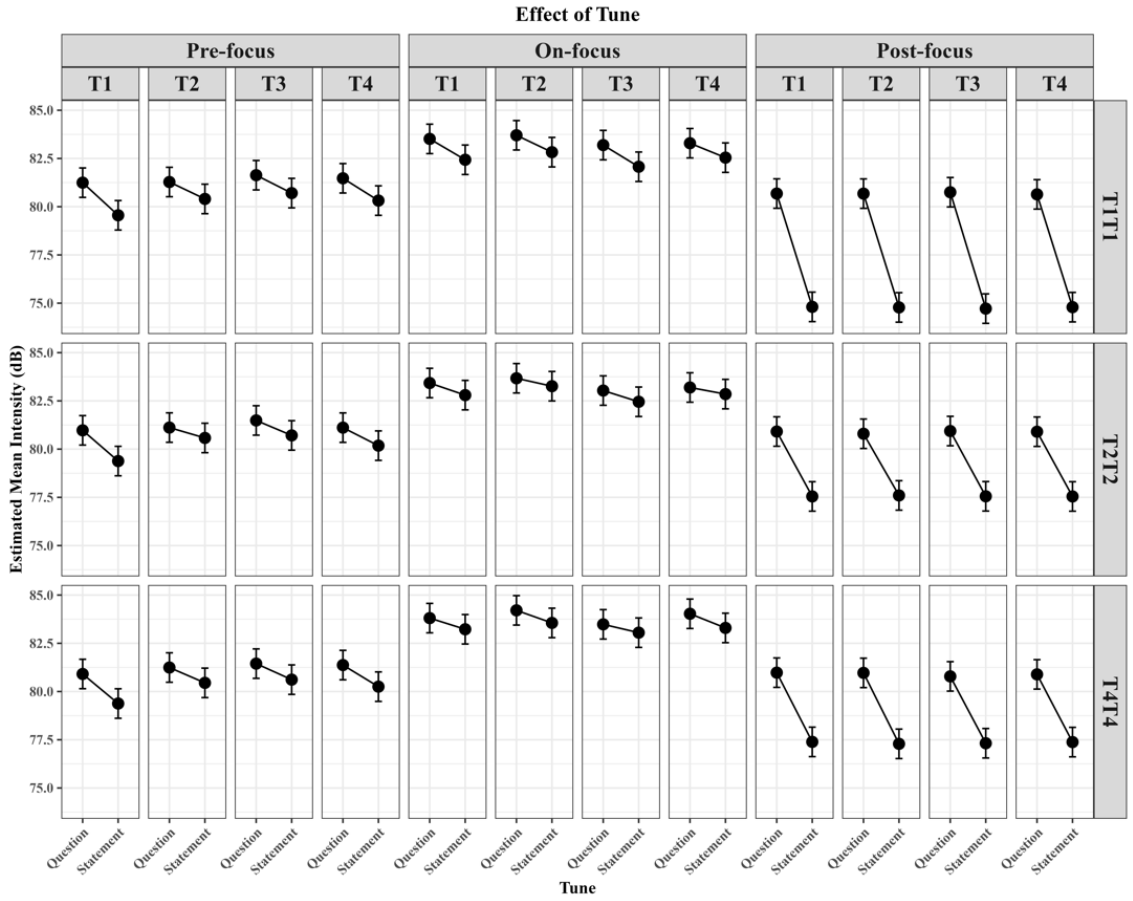


Figure 4.13 Effect of Tune on Mean Intensity across pre-target tones: Estimated marginal means with 95% confidence intervals

The Phrase also had a strong influence on intensity, as indicated by the significant main effect of Phrase ($F(2, 21{,}471.98) = 16{,}892.90$, $p < .001$). Intensity across phrase positions followed a pattern in which on-focus > pre-focus > post-focus. The on-focus phrases were realised with the highest intensity, consistent with canonical prosodic focus marking. Pre-focus intensity was intermediate, and post-focus phrases showed the lowest intensity overall. For example, when the focus was the T4T4 with T4 preceded, for the questions, the on-focus region was

significantly more intense than the pre-focus region, β = -2.66, t(21472) = -22.80, p < .0001. The difference between the on-focus and post-focus regions was also highly significant, with on-focus exceeding post-focus intensity by β = 3.14, t(21472) = 26.94, p < .0001, suggesting a gradual decline in intensity towards the end of the utterance. The pre-focus region was more intense than the post-focus region, β = 0.48, t(21472) = 4.14, p = .0001. In the case of statements, on-focus intensity was significantly greater than pre-focus, β = -3.04, t(21472) = -24.42, p < .0001 and post-focus, β = 5.92, t(21472) = 47.46, p < .0001. Pre-focus was significantly more intense than post-focus, β = 2.87, t(21472) = 23.04, p < .0001. Figure 4.14 shows the estimated marginal means of mean intensity across different tunes for each tone, separated by pre-target tones.

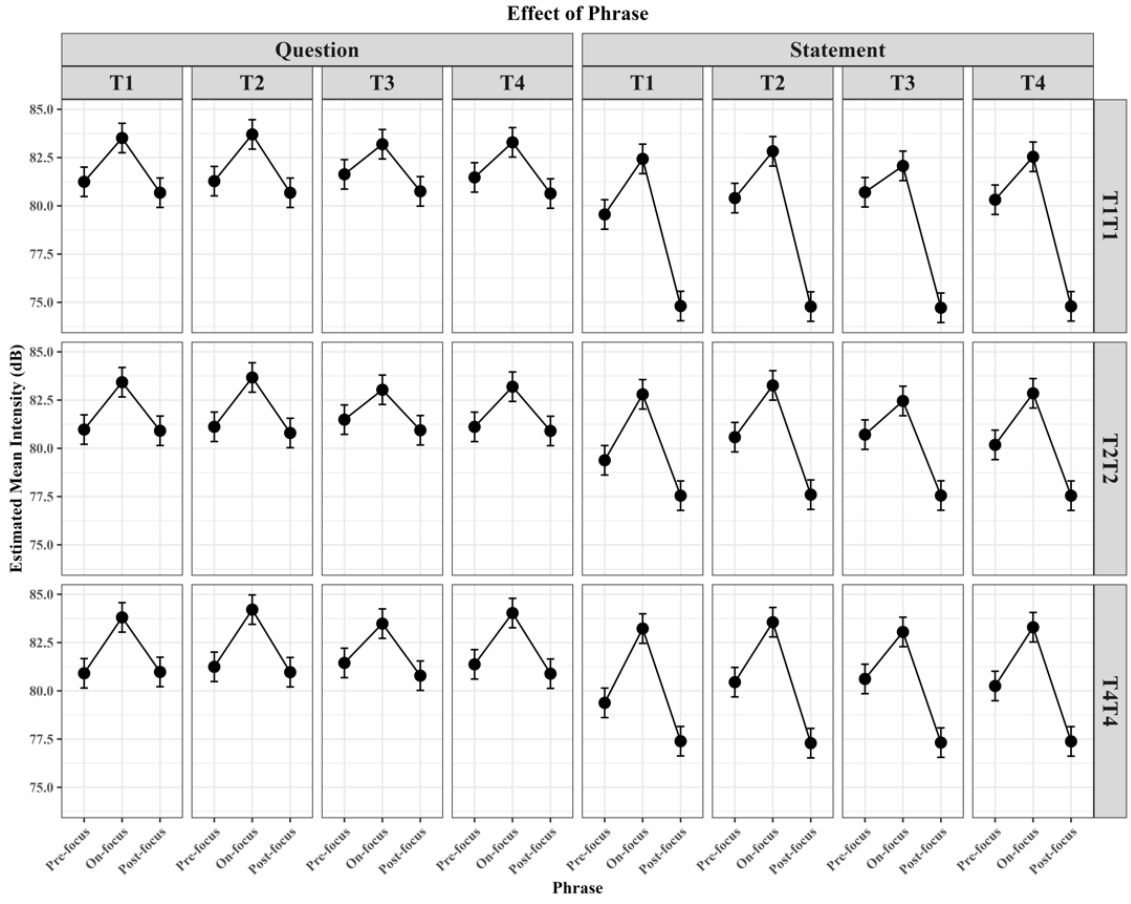


Figure 4.14 Effect of Phrase on Mean Intensity across pre-target tones: Estimated marginal means with 95% confidence intervals

The influence of pre-target tone on intensity was also significant (pre-target tone: (F(3, 21,473.30) = 38.15, p < .001) and significant in the interaction with Phrase (F(6, 21,471.98) = 65.04, p < .001). In the on-focus region, for instance, the T1 pre-target tone (low falling tone) was associated with relatively large on-focus contrasts. Under T1T1, the question-statement difference in on-focus intensity was β = 1.08 (t(21472) = 8.96, p < .0001), whereas under T4 pre-target tone (a high-level tone), it was β = 0.75 (t(21473) = 6.18, p < .0001). In the pre-focus region, for example, under T2T2 with T1 pre-target tone, the pre-focus contrast between questions and statements was β = 1.59 (t(21473) = 13.32, p < .0001), while the same tone with T2 pre-target tone yielded a smaller contrast of β = 0.54 (t(21472) = 4.47, p < .0001). See Table A.1 in Appendix A for the detailed mean intensity results.

4.4.4.2 Syllable-Level (On-Focus)

In the on-focus phrase (as shown in Table 4.4), a significant main effect of Tune was found (F(1, 14,343.40) = 548.30, p < .001), indicating that questions were produced with greater mean intensity than statements. Additionally, Tone (F(2, 44.12) = 8.52, p < .001), Syllable (F(1, 14,341.98) = 11.40, p < .001), and Pre-target (F(3, 14,342.52) = 75.15, p < .001) all had significant effects on intensity. The interaction between Tone and Syllable was particularly strong (F(2, 14,341.98) = 102.45, p < .001), as was the three-way interaction of Tune × Tone × Syllable (F(2, 14,341.98) = 44.98, p < .001).

Table 4.4 Effects of Tune, Tone, Syllable, and Pre-target on Mean Intensity

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	1,698.76	1,698.76	1	14,343.40	548.30	< .001***
Tone	52.79	26.39	2	44.12	8.52	< .001***
Syllable	35.31	35.31	1	14,341.98	11.40	< .001***
Pre-target	698.46	232.82	3	14,342.52	75.15	< .001***
Tune:Tone	248.21	124.10	2	14,342.69	40.06	< .001***
Tune:Syllable	2.76	2.76	1	14,341.98	0.89	.3450
Tone:Syllable	634.83	317.41	2	14,341.98	102.45	< .001***
Tune:Pre-target	12.03	4.01	3	14,342.64	1.29	.2743
Tone:Pre-target	26.99	4.50	6	14,342.51	1.45	.1905
Syllable:Pre-target	464.75	154.92	3	14,341.98	50.00	< .001***
Tune:Tone:Syllable	278.75	139.37	2	14,341.98	44.98	< .001***
Tune:Tone:Pre-target	50.02	8.34	6	14,342.64	2.69	.0130*
Tune:Syllable:Pre-target	7.43	2.48	3	14,341.98	0.80	.4938
Tone:Syllable:Pre-target	24.28	4.05	6	14,341.98	1.31	.2503
Tune:Tone:Syllable:Pre-target	21.11	3.52	6	14,341.98	1.14	.3385

The post-hoc comparisons showed that questions had higher mean intensity than statements, though the size of this effect varied by syllable and tonal context. For example, under Tone = T1T1, pre-target tone = T1, Syllable 1 in questions was significantly more intense than in statements by 0.78 dB (t(14,342) = 5.43, p < .0001), and Syllable 2 by an even greater 1.54 dB (t(14,342) = 10.66, p < .0001). A similar pattern held for Tone = T2T2, pre-target tone = T1, where the question-statement contrast for Syllable 1 was 1.02 dB (t(14,342) = 7.16, p < .0001), though the effect on Syllable 2 was not significant (estimate = 0.14, t(14,342) = 0.96, p = .34).

The intensity differences between Syllable 1 and Syllable 2 were not consistent in significance. In some cases, Syllable 1 was significantly more intense, while in others, Syllable 2 was either more intense or the difference was not significant. For instance, under Tone = T1T1, pre-target tone = T2, Tune = Statement, Syllable 1 was significantly more intense than Syllable 2, with an estimated difference of 1.46 dB (t(14,342) = 9.65, p < .0001). However, under Tone = T2T2, pre-target tone = T3, Tune = Statement, Syllable 1 was less intense than Syllable 2 by -1.06 dB (t(14,342) = -7.25, p < .0001). In other cases, such as T1T1, pre-target tone = T1, Tune =

Question, the difference was non-significant (estimate = 0.14, $t(14,342) = 1.01$, $p = .32$). The detailed mean intensity across tone, syllable and pre-target tone is listed in Table A.2 in Appendix A.

4.4.5 Mean F0

4.4.5.1 Phrase-Level

Significant main effect of Tune was found for mean F0 ($F(1, 21,481.08) = 11,347.23$, $p < .001$) (see Table 4.5). Across all tones, pre-target tones and phrases, mean F0 of questions was significantly higher than statements, suggesting a higher pitch register. The difference between questions and statements was especially large in the post-focus phrase, where the difference was as high as $\beta = 93.60$, $t(21475) = 32.80$, $p < .0001$ (e.g., when on-focus was T1T1, pre-target tone was T4), indicating a steep rise in pitch toward the end of interrogative utterances. Elevated F0 in questions was also observed in the pre-focus and on-focus phrases, though to a lesser extent, which confirms that the questions have a global higher register, not locally to the last syllable of the utterances.

Table 4.5 Effects of Tune, Tone, Phrase, and Pre-target on Mean F0

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	13,643,007.24	13,643,007.24	1	21,481.08	11,347.23	< .001***
Tone	559,412.45	279,706.22	2	41.90	232.64	< .001***
Phrase	2,564,266.66	1,282,133.33	2	21,471.84	1,066.38	< .001***
Pre-target	245,548.04	81,849.35	3	21,478.01	68.08	< .001***
Tune:Tone	50,472.12	25,236.06	2	21,479.96	20.99	< .001***
Tune:Phrase	1,841,764.35	920,882.17	2	21,471.84	765.92	< .001***
Tone:Phrase	1,815,772.17	453,943.04	4	21,471.84	377.56	< .001***
Tune:Pre-target	11,378.32	3,792.77	3	21,479.47	3.15	.0237*
Tone:Pre-target	4,519.52	753.25	6	21,478.00	0.63	.7092
Phrase:Pre-target	303,283.88	50,547.31	6	21,471.84	42.04	< .001***
Tune:Tone:Phrase	181,623.76	45,405.94	4	21,471.84	37.77	< .001***
Tune:Tone:Pre-target	6,102.02	1,017.00	6	21,479.43	0.85	.5342
Tune:Phrase:Pre-target	11,873.36	1,978.89	6	21,471.84	1.65	.1301

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tone:Phrase:Pre-target	15,207.80	1,267.32	12	21,471.84	1.05	.3951
Tune:Tone:Phrase:Pre-target	10,173.28	847.77	12	21,471.84	0.71	.7481

In a further examination of the effects of tone, sentence type and phrase position and their interactions on mean F0, post-hoc comparisons were conducted using estimated marginal means averaged over the levels of pre-target tone, since pre-target tone interacts with both Tone and Phrase, as supported by a significant Phrase × pre-target tone interaction in Table 4.5), and its effects are addressed separately in the following part. By averaging across pre-target tones, it is easier to examine mean F0 patterns which are shaped by sentence modality (question vs. statement), lexical tone (T1T1, T2T2, T4T4), and prosodic phrasing (pre-focus, on-focus, post-focus).

Across all tone types and phrase positions, questions were significantly higher in mean F0 than statements (all *p*s < .0001). The question-statement contrast was the largest in the post-focus position across all tones, with particularly large differences observed in T1T1 (91.10 Hz, $p < .0001$), followed by T4T4 (69.50 Hz, $p < .0001$) and T2T2 (67.50 Hz, $p < .0001$), which suggests a strong high boundary tone effect specific to questions that elevates pitch considerably in final positions.

For T1T1, the pitch contour was relatively flat in questions, with no significant difference between the on-focus and post-focus regions ($\beta = 2.10$, $t(21472) = 1.52$, $p = .28$), and only a modest increase from pre-focus to on-focus ($\beta = 4.12$, $t(21472) = 2.98$, $p = .01$). In contrast, T1T1 in statements maintained stable pitch from pre-focus to on-focus ($\beta = 1.48$, $p = .57$) but showed a dramatic drop in the post-focus region ($\beta = 53.93$, $p < .0001$). This pattern indicates that question intonation flattens the pitch in T1T1 in the on-focus phrase with the global higher

register and the pitch height of the post-focus phrases by the high boundary tone, while statements preserve a post-focus F0 fall.

In questions, post-focus pitch for T2T2 actually exceeded on-focus pitch (β = -3.93, p = .01), indicating a strong rising boundary tone effect specific to questions. Both pre-focus to on-focus transitions showed significant increases in questions (β = -14.39, p < .0001), while in statements, T2T2 maintained similar pitch levels from pre-focus to on-focus (β = 0.96, p = .78) before dropping significantly in the post-focus position (β = 15.35, p < .0001).

For T4T4, both questions and statements showed that the mean F0 of the on-focus phrases was the highest, with significant increases from pre-focus to on-focus positions (questions: β = -47.43, p < .0001; statements: β = -38.52, p < .0001), followed by significant decreases in the post-focus region (questions: β = 30.97, p < .0001; statements: β = 59.77, p < .0001). However, the post-focus drop was substantially larger in statements than questions, resulting in a 69.50 Hz question-statement difference in the post-focus position compared to 40.70 Hz at the on-focus site. See Figure 4.15 below for the estimated marginal means of mean F0 with 95% confidence intervals, illustrating the effect of Tune (left panel) and Phrase (right panel) across tones. The detailed averaged mean F0 across tone, phrase and pre-target tone is listed in Table A.1 in Appendix A.

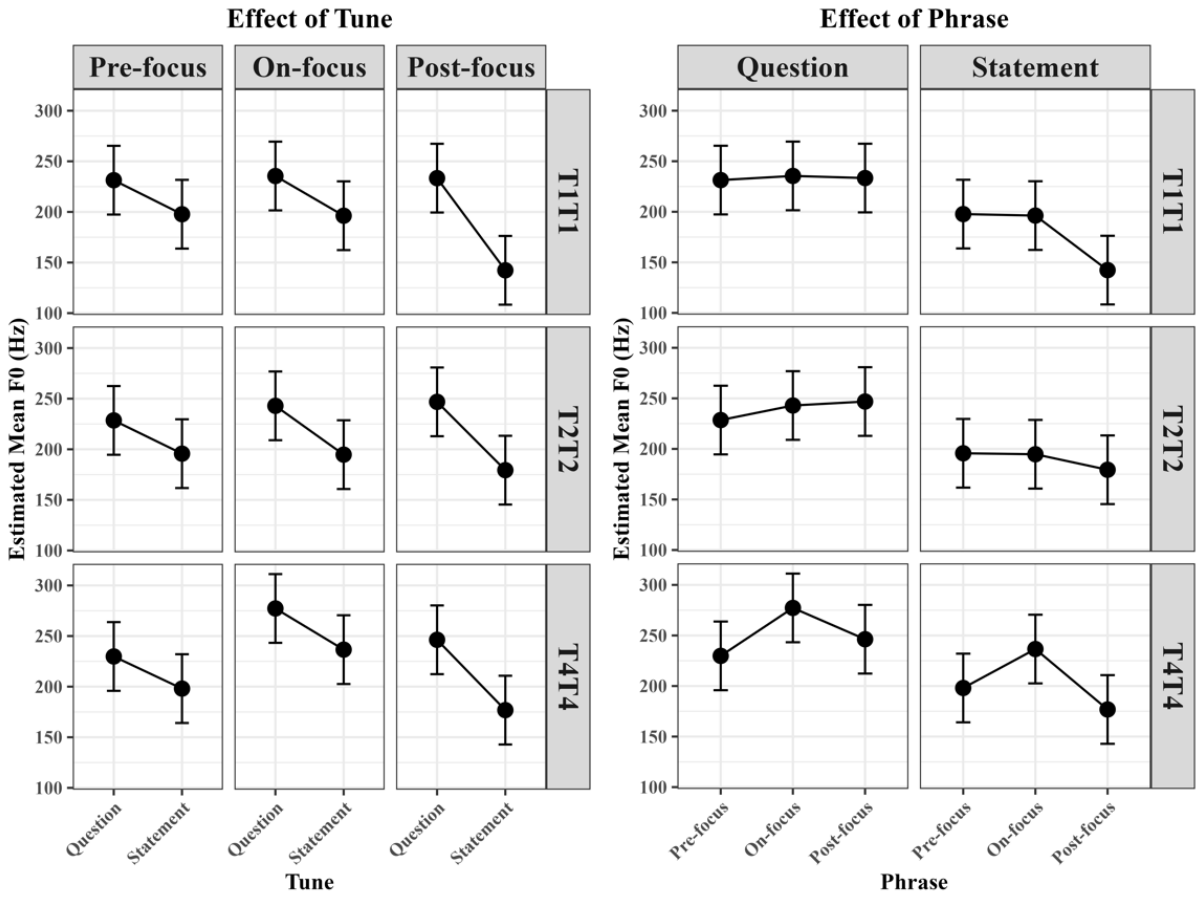


Figure 4.15 Effects of Tune (left panel) and Phrase (right panel) on Mean F0: Estimated marginal means with 95% confidence intervals

With the pre-target tone effect included, Figure 4.16 shows that questions consistently display higher mean F0 than statements across all conditions, with the difference being especially pronounced in the post-focus position. In the post-focus position, the question-statement differences range from approximately 66 Hz to 93 Hz (all $p < .0001$), which is substantially larger than in the pre-focus (22-37 Hz) or on-focus positions (38-50 Hz). For example, with pre-target tone T1 and T1T1, the question-statement contrast in post-focus position was 88.00 Hz ($t(21474) = 30.93$, $p < .0001$), compared to 30.30 Hz in the pre-focus and 38.90 Hz in the on-focus positions. This pattern of dramatically increased question-statement contrast in the post-focus position is consistent across all tone types and pre-target tones, though with some

variation in magnitude. T1T1 sequences show the largest post-focus question-statement differences (ranging from 88.00 to 93.60 Hz), while T2T2 and T4T4 show somewhat smaller differences (ranging from 66.00 to 74.50 Hz).

The influence of the pre-target tone on mean F0 patterns was evident in its modulation of the F0 mean of these phrasal contrasts. In the case of T1T1, pre-tonal influence created distinctly different patterns across phrasal positions. Consideration of pre-target tone T3 (high falling tone) in the statement tune revealed that the pre-focus tone exhibited a significantly higher mean F0 than the on-focus tone (β = 9.50, t(21472) = 3.26, p = .003), while pre-target tone T1 (low falling tone) and the pre-focus tone exhibited lower mean F0 values than the on-focus tone (β = -9.94, t(21472) = -3.39, p = .002). For T2T2, pre-tonal influence was most visible in the question tune, where the magnitude of the pre-focus to on-focus contrast ranged from -29.47 Hz to just -5.70 Hz for pre-target tone T1 and pre-target tone T4, respectively. For T4T4, the on-focus-post contrast in the statement tune showed substantial pre-tonal influence, ranging from 53.89 Hz to 68.24 Hz for pre-target tone T1 and pre-target tone T4, respectively. This suggests that even for a high-level tone, the preceding tonal context significantly influences how mean F0 was distributed across the utterance.

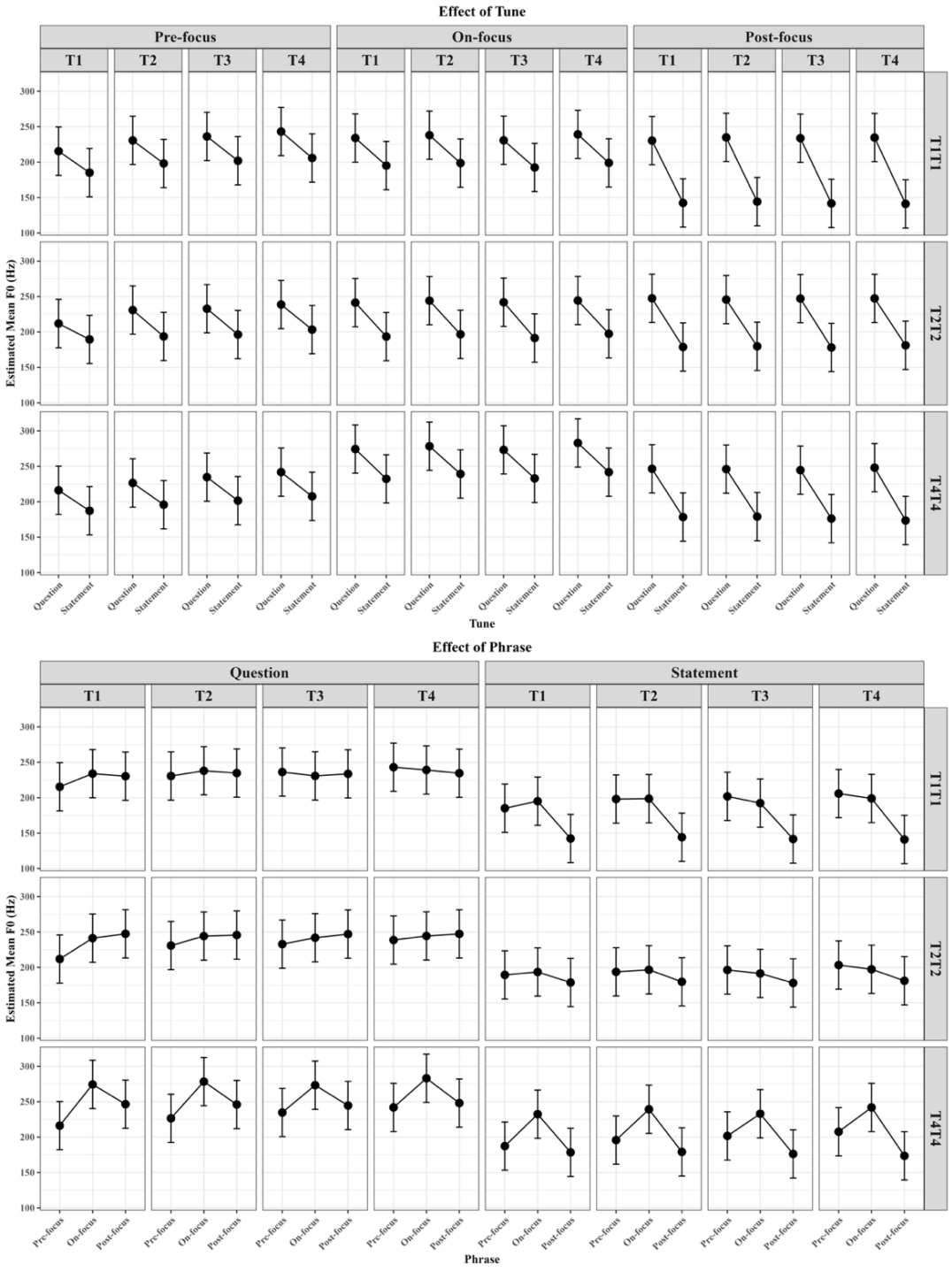


Figure 4.16 Effects of Tune (top panel) and Phrase (bottom panel) on Mean F0 across pre-target tones: Estimated marginal means with 95% confidence intervals

4.4.5.2 Syllable-Level

The main effect of Tune was significant (F(1, 14,344.01) = 6,734.80, p < .001), confirming that questions are produced with significantly higher F0 than statements. There were also significant effects of Tone (F(2, 43.97) = 203.59, p < .001), Syllable (F(1, 14,341.95) = 1,719.94, p < .001), and Pre-target tone (F(3, 14,343.01) = 43.64, p < .001). Interactions between these factors, including Tune × Tone, Tune × Syllable, Tone × Syllable, and Tune × Tone × Syllable, were all significant (all p < .001), as shown in Table 4.6.

Table 4.6 Effects of Tune, Tone, Syllable, and Pre-target on Mean F0

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	6,982,515.18	6,982,515.18	1	14,344.01	6,734.80	< .001***
Tone	422,154.53	211,077.26	2	43.97	203.59	< .001***
Syllable	1,783,207.39	1,783,207.39	1	14,341.95	1,719.94	< .001***
Pre-target	135,722.55	45,240.85	3	14,343.01	43.64	< .001***
Tune:Tone	48,379.07	24,189.54	2	14,343.35	23.33	< .001***
Tune:Syllable	22,580.65	22,580.65	1	14,341.95	21.78	< .001***
Tone:Syllable	309,122.72	154,561.36	2	14,341.95	149.08	< .001***
Tune:Pre-target	1,257.41	419.14	3	14,343.24	0.40	.7499
Tone:Pre-target	16,991.04	2,831.84	6	14,342.99	2.73	.0118*
Syllable:Pre-target	159,010.22	53,003.41	3	14,341.95	51.12	< .001***
Tune:Tone:Syllable	19,493.90	9,746.95	2	14,341.95	9.40	< .001***
Tune:Tone:Pre-target	3,697.05	616.18	6	14,343.24	0.59	.7352
Tune:Syllable:Pre-target	1,106.13	368.71	3	14,341.95	0.36	.7851
Tone:Syllable:Pre-target	7,733.63	1,288.94	6	14,341.95	1.24	.2805
Tune:Tone:Syllable:Pre-target	1,625.90	270.98	6	14,341.95	0.26	.9548

Similarly, when averaged across pre-target tone levels, questions consistently showed higher mean F0 than statements across all tones and syllables. For instance, in T1T1, the mean F0 difference between question and statement was 40.10 Hz for Syllable 1 (t(14,342) = 30.25, p < .0001) and 41.90 Hz for Syllable 2 (t(14,342) = 31.63, p < .0001). In T2T2, these differences increased to 43.60 Hz (Syllable 1, t(14,344) = 33.22, p < .0001) and 55.20 Hz (Syllable 2, t(14,344) = 42.05, p < .0001). Similar effects were found in T4T4, with 41.50 Hz and 43.10 Hz differences for Syllables 1 and 2, respectively (both p < .0001).

Furthermore, averaged over pre-target tone, Syllable 2 always had significantly higher mean F0 than Syllable 1, suggesting that the prominence was in Syllable 2 in the on-focus phrases. For example, in T2T2, the contrast between Syllable 1 and Syllable 2 was -39.00 Hz in questions (t(14,342) = -30.42, p < .0001) and -27.40 Hz in statements (t(14,342) = -20.50, p < .0001). Even in T1T1, where tone contours were flatter in questions, Syllable 2 still had a significantly higher F0: -11.40 Hz in questions (t(14,342) = -8.91, p < .0001) and -9.60 Hz in statements (t(14,342) = -7.03, p < .0001). The pattern held for T4T4 as well (-23.90 Hz in questions, -22.30 Hz in statements, both p < .0001).

As for the pre-target tone effect, the results indicate that the pre-target tone had a significant influence on mean F0, particularly on Syllable 1. Statistically, this was supported by a significant Syllable × pre-target tone interaction (F(3, 14,341.95) = 51.12, p < .001), confirming that the effect of pre-target tone differs across syllable positions (as shown in Table 4.6). Although Syllable 2 consistently showed a high F0 and was relatively stable across tonal contexts, Syllable 1 exhibited greater variability in F0 depending on the tonal shape of the preceding syllable. For example, in the T1T1 tone condition under question intonation, the Syllable 1-Syllable 2 contrast was -17.71 Hz following a T1 pre-target tone, but reduced to -4.26 Hz following T2, and increased again with T3 (-16.13 Hz), demonstrating the modulating effect of pre-target tone. This pattern reflects coarticulatory effects, whereby the F0 contour of the preceding tone affects the pitch target of the immediately following syllable. As Syllable 2 is more prosodically autonomous, often marking phrase boundaries or carrying nuclear pitch accents, it appears to resist the influence of pre-target tone, resulting in more stable F0 realisation.

For the T1T1 sequence, with question intonation, the inter-syllabic F0 difference varies dramatically based on pre-target tone. When preceded by T2 (rising tone), the syllable difference was merely -4.26 Hz ($t(14,342) = -1.66$, $p = .10$) and not statistically significant, suggesting that the rising endpoint of pre-target tone T2 aligns well with the rising component of T1T1, creating a smooth tonal transition that minimises the pitch difference between syllables. In contrast, when T1T1 is preceded by T1 (low falling tone), the difference increases to -17.71 Hz ($t(14,342) = -6.92$, $p < .0001$), indicating that the low endpoint of the preceding T1 creates a lower starting point for the first syllable, amplifying the inter-syllabic contrast.

For the T2T2 sequence, the largest syllable differences (-51.27 Hz) were observed when preceded by T1 (low falling tone). This pattern makes phonetic sense because the low endpoint of pre-target tone T1 establishes a particularly low starting point for the first rising syllable of T2T2, resulting in a more pronounced rise across the disyllabic unit. When preceded by T2 (another rising tone), the difference decreases to -29.44 Hz, likely because the high endpoint of the pre-target tone T2 raises the starting point of the following T2, reducing the overall rising excursion.

For the T4T4 sequence, the pre-target tone effects are less dramatic but still significant. When preceded by T3 (high falling tone), the syllable difference is -25.83 Hz, compared to -17.22 Hz when preceded by T2 (rising tone). The relative stability of T4T4 across different pre-tonal environments, compared to T1T1 and T2T2, aligns with its phonetic identity as a high-level tone, which may be less susceptible to coarticulatory perturbation than contour tones. The detailed averaged mean F0 across tone, syllable and pre-target tone is listed in Table A.2 in Appendix A.

4.4.6 Maximum F0

4.4.6.1 Phrase-Level

The results revealed significant main effects of Tune ($F(1, 21{,}482.21) = 3{,}569.95$, $p < .001$), Tone ($F(2, 42.02) = 109.23$, $p < .001$), and Phrase ($F(2, 21{,}471.96) = 121.66$, $p < .001$). The main effect of pre-target tone was not significant ($F(3, 21{,}478.94) = 1.17$, $p = .32$). Several significant interactions were observed: Tune × Tone ($F(2, 21{,}481.12) = 21.22$, $p < .001$), Tune × Phrase ($F(2, 21{,}471.96) = 154.09$, $p < .001$), Tone × Phrase ($F(4, 21{,}471.96) = 215.88$, $p < .001$), and a three-way interaction of Tune × Tone × Phrase ($F(4, 21{,}471.96) = 37.91$, $p < .001$). Significant but smaller interactions involving pre-target tone were also found: Tune × pre-target tone ($F(3, 21{,}480.59) = 3.61$, $p = .01$), Phrase × pre-target tone ($F(6, 21{,}471.96) = 2.80$, $p = .01$), and Tune × Phrase × pre-target tone ($F(6, 21{,}471.96) = 2.48$, $p = .02$). See Table 4.7.

Table 4.7 Effects of Tune, Tone, Phrase, and Pre-target on Maximum F0

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	11,469,586.36	11,469,586.36	1	21,482.21	3,569.95	< .001***
Tone	701,860.70	350,930.35	2	42.02	109.23	< .001***
Phrase	781,748.67	390,874.33	2	21,471.96	121.66	< .001***
Pre-target	11,296.08	3,765.36	3	21,478.94	1.17	.3187
Tune:Tone	136,347.08	68,173.54	2	21,481.12	21.22	< .001***
Tune:Phrase	990,100.95	495,050.48	2	21,471.96	154.09	< .001***
Tone:Phrase	2,774,357.54	693,589.38	4	21,471.96	215.88	< .001***
Tune:Pre-target	34,837.48	11,612.49	3	21,480.59	3.61	.0126*
Tone:Pre-target	16,505.58	2,750.93	6	21,478.94	0.86	.5263
Phrase:Pre-target	53,912.16	8,985.36	6	21,471.96	2.80	.0101*
Tune:Tone:Phrase	487,168.05	121,792.01	4	21,471.96	37.91	< .001***
Tune:Tone:Pre-target	16,232.78	2,705.46	6	21,480.54	0.84	.5371
Tune:Phrase:Pre-target	47,730.35	7,955.06	6	21,471.96	2.48	.0214*
Tone:Phrase:Pre-target	27,399.64	2,283.30	12	21,471.96	0.71	.7426
Tune:Tone:Phrase:Pre-target	63,978.96	5,331.58	12	21,471.96	1.66	.0688

When comparing question and statement tunes, questions consistently showed significantly higher maximum F0 than statements across all tone types and phrase positions. For T1T1, this question-statement difference was most dramatic in the post-focus position ($\beta = 90.10$, $t(21473) = 38.60$, $p < .0001$), compared to pre-focus ($\beta = 33.30$, $t(21473) = 14.28$, $p < .0001$) and focus positions ($\beta = 35.50$, $t(21473) = 15.22$, $p < .0001$). For T2T2, the question-statement difference increased progressively from pre-focus ($\beta = 33.30$, $t(21479) = 14.41$, $p < .0001$) to focus ($\beta = 48.80$, $t(21479) = 21.12$, $p < .0001$) to post-focus positions ($\beta = 53.40$, $t(21479) = 23.13$, $p < .0001$). Similarly, for T4T4, the question-statement difference was largest in the post-focus position ($\beta = 51.20$, $t(21474) = 22.14$, $p < .0001$), compared to pre-focus ($\beta = 32.20$, $t(21474) = 13.93$, $p < .0001$) and focus positions ($\beta = 38.90$, $t(21474) = 16.80$, $p < .0001$).

For T1T1 in questions, the focus position showed significantly higher maximum F0 than the pre-focus position ($\beta = -14.49$, $t(21472) = -6.41$, $p < .0001$). The post-focus position also displayed higher maximum F0 than the pre-focus position, though this difference was not statistically significant ($\beta = -3.04$, $t(21472) = -1.35$, $p = .37$). Specifically, the focus position exhibited significantly higher maximum F0 than the post-focus position ($\beta = 11.45$, $t(21472) = 5.07$, $p < .0001$). In statement contexts for the case of T1T1, although the focus position again showed higher maximum F0 than the pre-focus position ($\beta = -12.28$, $t(21472) = -5.11$, $p < .0001$), the post-focus position displayed a dramatically lower maximum F0 than both the pre-focus ($\beta = 53.71$, $t(21472) = 22.34$, $p < .0001$) and focus positions ($\beta = 65.99$, $t(21472) = 27.45$, $p < .0001$).

For T2T2 in questions, Maximum F0 increased progressively across phrase positions, with focus higher than pre-focus ($\beta = -13.99$, $t(21472) = -6.20$, $p < .0001$), post-focus higher than pre-focus ($\beta = -31.62$, $t(21472) = -14.00$, $p < .0001$), and post-focus higher than focus ($\beta = -$

17.63, $t(21472) = -7.81$, $p < .0001$). In statement contexts, the pattern differed, with no significant difference between pre-focus and focus positions ($\beta = 1.51$, $t(21472) = 0.64$, $p = .80$), but with post-focus showing higher Maximum F0 than both pre-focus ($\beta = -11.49$, $t(21472) = -4.88$, $p < .0001$) and focus positions ($\beta = -13.00$, $t(21472) = -5.52$, $p < .0001$).

For T4T4 in questions, Maximum F0 also increased progressively, with focus higher than pre-focus ($\beta = -27.88$, $t(21472) = -12.38$, $p < .0001$), post-focus higher than pre-focus ($\beta = -37.82$, $t(21472) = -16.79$, $p < .0001$), and post-focus higher than focus ($\beta = -9.94$, $t(21472) = -4.41$, $p < .0001$). In statement contexts, both focus and post-focus showed higher Maximum F0 than pre-focus ($\beta = -21.24$, $t(21472) = -8.96$, $p < .0001$ and $\beta = -18.84$, $t(21472) = -7.94$, $p < .0001$, respectively), but with no significant difference between focus and post-focus positions ($\beta = 2.40$, $t(21472) = 1.01$, $p = .57$). See Figure 4.17 for the effects of tune (left panel) and phrase (right panel) averaged by pre-target tones.

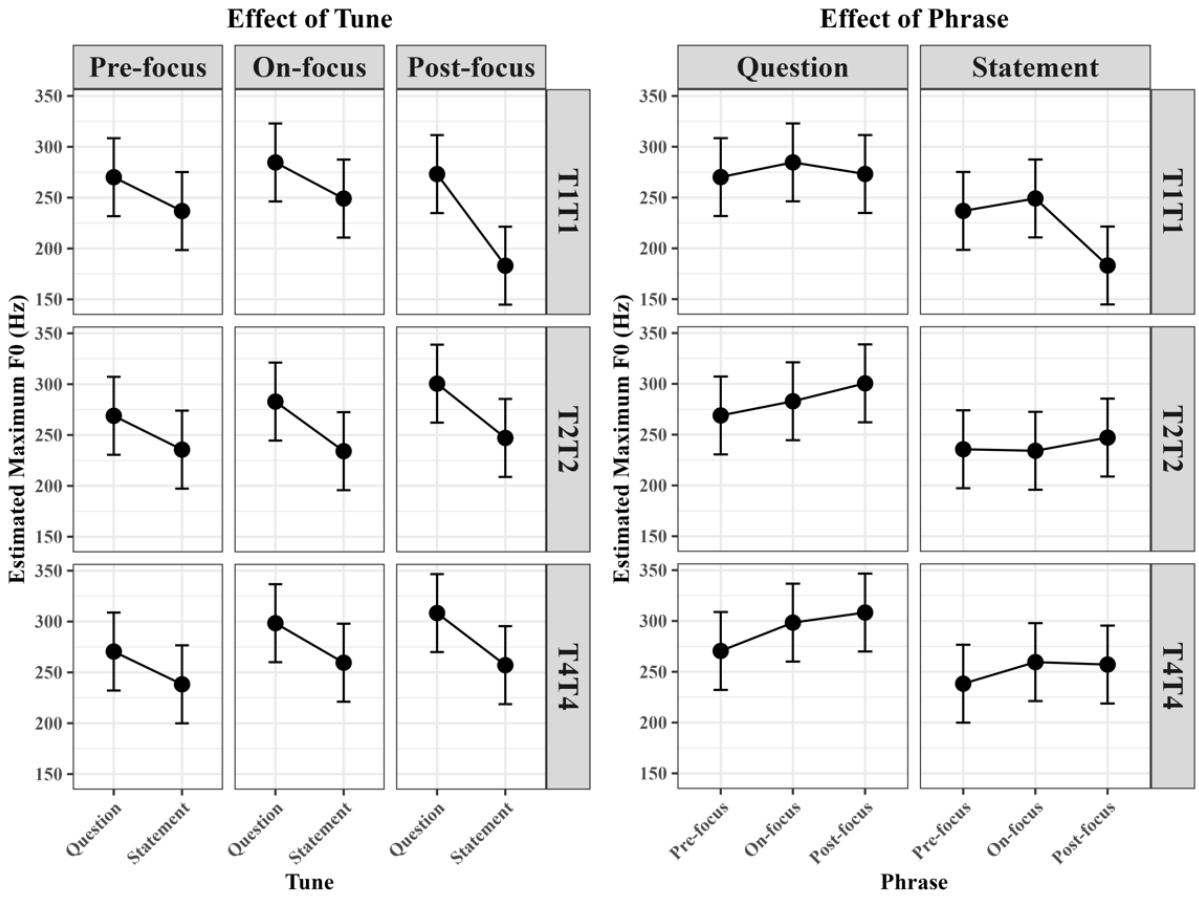


Figure 4.17 Effects of Tune (left panel) and Phrase (right panel) on Maximum F0: Estimated marginal means with 95% confidence intervals

With the pre-target tone factor included, Figure 4.18 still shows that questions consistently exhibit higher maximum F_0 than statements across all conditions, although the magnitude of this contrast varies across phrasal positions. Unlike the minimum F0 and mean F0 data, where post-focus position showed dramatically larger question-statement differences, the maximum F0 data shows more complex patterns.

For T1T1, the question-statement contrast is substantially larger in the post-focus position (ranging from 85.00 to 92.60 Hz) than in the pre-focus (27.20 to 42.60 Hz) or focus positions (30.50 to 40.20 Hz). For example, with the pre-target tone T2, the question-statement contrast

in post-focus position was 92.60 Hz (t(21476) = 19.71, p < .0001), compared to 30.30 Hz in pre-focus and 36.60 Hz in the focus positions. However, for T2T2 and T4T4, the question-statement differences are more evenly distributed across phrasal positions, particularly in the focus and post-focus positions. For T2T2 with pre-target tone T3, the question-statement contrast was 34.30 Hz in pre-focus, 53.00 Hz in focus, and 55.80 Hz in post-focus position. This suggests that the intonational effect on maximum F0 is more tone-dependent than was observed for minimum and mean F0.

The comparisons across phrasal positions revealed distinctly different patterns in questions versus statements. In the question tune, pre-focus position generally had lower maximum F0 than both focus and post-focus positions across most tonal combinations. For instance, with T4T4 and pre-target tone T1, pre-focus had significantly lower maximum F0 than focus (β = -30.52, t(21472) = -6.79, p < .0001) and post-focus (β = -39.39, t(21472) = -8.76, p < .0001). In the statement tune, the pattern varied substantially by tone type. For T1T1, pre-focus and focus positions had similar maximum F0, but the post-focus position showed a dramatically lower maximum F0. With pre-target tone T3, the pre-post contrast was 63.57 Hz (t(21472) = 13.34, p < .0001) and the focus-post contrast was 68.04 Hz (t(21472) = 14.28, p < .0001). For T2T2 and T4T4, the differences across phrasal positions were more modest, with smaller contrasts between pre-focus and post-focus positions.

For T1T1, pre-tonal influence created different patterns across phrasal positions in the question tune. With pre-target tone T3 (high falling tone), pre-focus and focus positions had similar maximum F0 (β = -1.78, t(21472) = -0.39, p = .92), while with pre-target tone T1 (low falling tone), pre-focus had significantly lower maximum F0 than focus (β = -21.65, t(21472) = -4.81, p < .0001). For T2T2, pre-tonal influence was particularly evident in pre-focus to focus

contrasts in the question tune, with values ranging from -25.75 Hz with pre-target tone T1 to just -4.16 Hz with pre-target tone T2 (non-significant, p = .63). For T4T4, the pre-focus to focus contrast in the question tune showed less pre-tonal variation, ranging from -24.78 Hz to -30.79 Hz across different pre-target tones. See mean maximum F0 across tone, phrase and pre-target tone in Table A.1 in Appendix A.

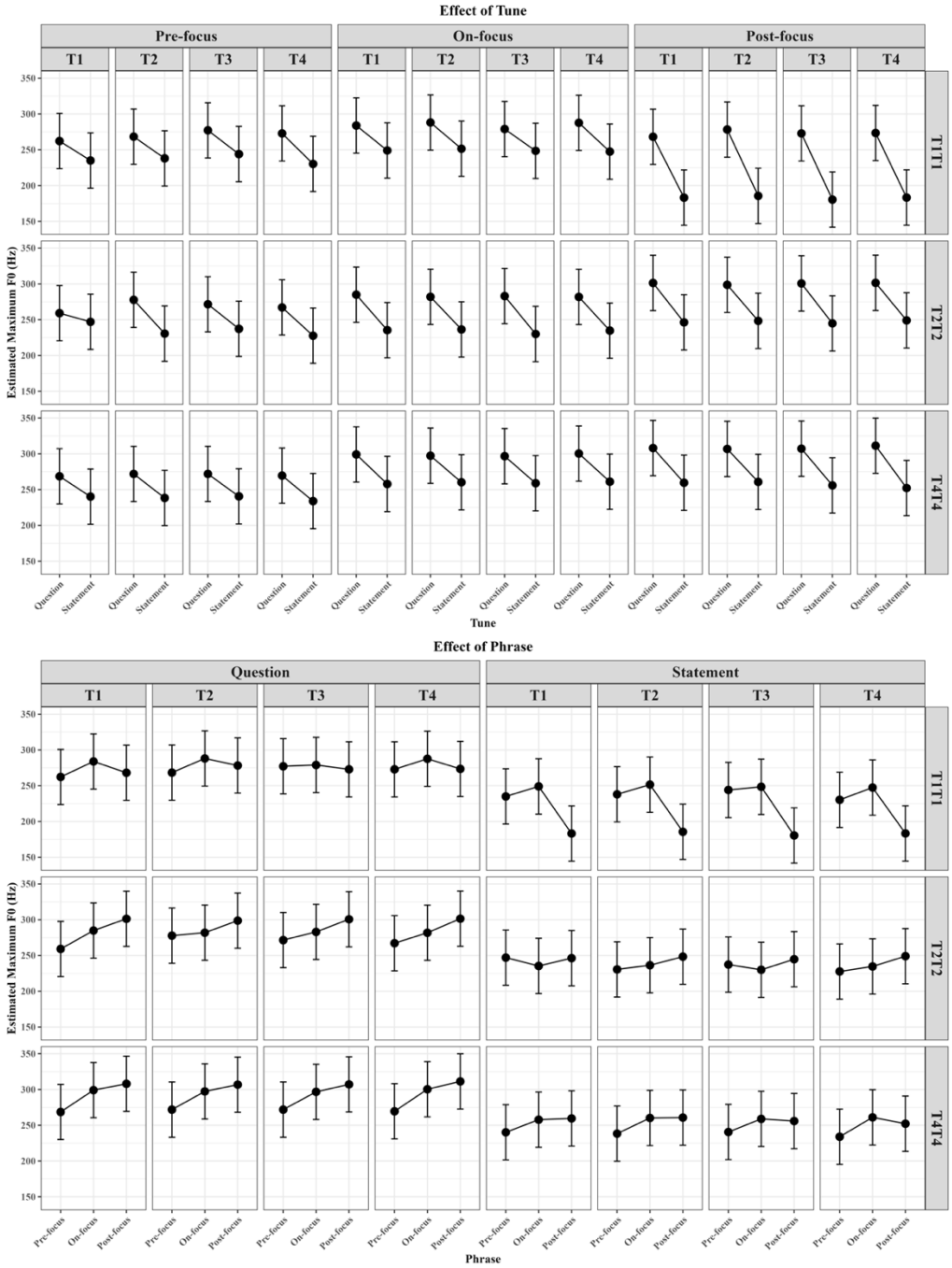


Figure 4.18 Effects of Tune (top panel) and Phrase (bottom panel) on Maximum F0 across pre-target tones: Estimated marginal means with 95% confidence intervals

4.4.6.2 Syllable-Level (On-Focus)

The statistical result revealed significant effects of Tune ($F(1, 14{,}343.23) = 4{,}714.36$, $p < .001$), Tone ($F(2, 43.98) = 28.07$, $p < .001$), Syllable ($F(1, 14{,}341.95) = 103.27$, $p < .001$), and pre-target tone ($F(3, 14{,}342.74) = 5.19$, $p = .001$). The interactions told a different story from mean F0. The Tune × Syllable interaction for maximum F0 was not significant ($F(1, 14{,}341.95) = 1.05$, $p = .30$), indicating that the effect of tune on maximum F0 was relatively consistent across syllable positions. The Tone × pre-target tone and Tune × pre-target tone interactions were likewise non-significant. See Table 4.8 for the detailed statistical results. See Table 4.8 for the full statistical results.

Table 4.8 Effects of Tune, Tone, Syllable, and Pre-target on Maximum F0

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	7,114,135.10	7,114,135.10	1	14,343.23	4,714.36	< .001***
Tone	84,729.24	42,364.62	2	43.98	28.07	< .001***
Syllable	155,841.94	155,841.94	1	14,341.95	103.27	< .001***
Pre-target	23,488.01	7,829.34	3	14,342.74	5.19	.0014**
Tune:Tone	65,089.31	32,544.65	2	14,343.00	21.57	< .001***
Tune:Syllable	1,589.40	1,589.40	1	14,341.95	1.05	.3048
Tone:Syllable	213,097.08	106,548.54	2	14,341.95	70.61	< .001***
Tune:Pre-target	2,145.60	715.20	3	14,342.93	0.47	.7004
Tone:Pre-target	15,063.75	2,510.63	6	14,342.73	1.66	.1255
Syllable:Pre-target	18,869.00	6,289.67	3	14,341.95	4.17	.0059**
Tune:Tone:Syllable	8,096.83	4,048.42	2	14,341.95	2.68	.0684
Tune:Tone:Pre-target	11,674.39	1,945.73	6	14,342.93	1.29	.2581
Tune:Syllable:Pre-target	633.61	211.20	3	14,341.95	0.14	.9361
Tone:Syllable:Pre-target	1,921.51	320.25	6	14,341.95	0.21	.9731
Tune:Tone:Syllable:Pre-target	984.13	164.02	6	14,341.95	0.11	.9955

When the post-hoc test was averaged by pre-target tone, regarding tune differences (i.e., question vs. statement), all conditions showed significantly higher maximum F0 in questions than in statements. For Syllable 1, the difference was β = 41.10 (1.60), t(14342) = 25.69, p < .0001 in T1T1, β = 47.90 (1.58), t(14343) = 30.24, p < .0001 in T2T2, and β = 43.00 (1.59), t(14342) = 27.13, p < .0001 in T4T4. For Syllable 2, the pattern remained consistent, with β = 40.80 (1.60), t(14342) = 25.53, p < .0001 in T1T1, β = 53.40 (1.58), t(14343) = 33.73, p < .0001 in T2T2, and β = 41.70 (1.59), t(14342) = 26.33, p < .0001 in T4T4.

As for the comparisons between the two syllables, for T1T1, the maximum F0 was significantly higher in Syllable 1 compared to Syllable 2, both in the questions, β = 3.71 (1.55), t(14342) = 2.40, p = .02, and the statements, β = 3.45 (1.65), t(14342) = 2.10, p = .04. In contrast, for T2T2, Syllable 2 exhibited significantly higher maximum F0 than Syllable 1, both in questions, β = -17.88 (1.55), t(14342) = -11.55, p < .0001, and in statements, β = -12.35 (1.61), t(14342) = -7.65, p < .0001. A similar pattern was observed for T4T4, where Syllable 2 had greater maximum F0 than Syllable 1, both in questions, β = -7.60 (1.54), t(14342) = -4.92, p < .0001, and statements, β = -8.87 (1.63), t(14342) = -5.46, p < .0001.

With the inclusion of pre-target tone as a factor, the post-hoc results revealed that pre-target tone had more influence on Syllable 1 than on Syllable 2, in line with findings from the mean F0 analysis. As Syllable 1 immediately follows the pre-target tone, it is more susceptible to tonal coarticulation and carryover effects. For instance, in the T2T2 tonal environment with question intonation, the contrast between Syllable 1 and Syllable 2 varied depending on the pre-target tone: it was -21.01 Hz following T1 (low falling), -13.86 Hz following T2 (rising), -20.73 Hz following T3 (high falling), and -15.92 Hz following T4 (high-level). Low or falling

pre-target tones tend to preserve large pitch differences across the two syllables, while rising or high-level pre-target tones compress the pitch contour by raising the F0 of Syllable 1.

In comparison, Syllable 2 showed a consistent increase in maximum F0 regardless of pre-target tone, likely due to its role as a locus of prosodic prominence, often bearing the final pitch accent or boundary tone. This implies that Syllable 2 is less influenced by preceding tones, and more closely reflects the pitch target associated with the tone of the syllable itself or the overall sentence-level intonation. For example, under Tone = T2T2, pre-target tone = T1, the question-statement difference in maximum F0 for Syllable 2 was 53.60 Hz ($t = 16.99$, $p < .0001$), and it remained similarly high across all other pre-target tones (ranging from 51.00 to 57.10 Hz), emphasising its resistance to tonal carryover effects.

Interestingly, in tone sequences like T1T1 (rising-falling), the Syllable 1-Syllable 2 contrast in questions was frequently non-significant, regardless of pre-target tone. For example, when the pre-target tone = T1 or T3, the contrasts were just 1.48 Hz and 1.16 Hz, respectively (both $p > .60$). This suggests that the inherent shape of T1T1 tones, which involve both a pitch rise and a fall within the tone pair on the lexical tone level but with a global register raising by question intonation, may neutralise the influence of pre-target tone on peak F0, or result in overlapping pitch targets that obscure inter-syllabic contrasts. See mean maximum F0 across tone, syllable and pre-target tone in Table A.2 in Appendix A.

4.4.7 Minimum F0

4.4.7.1 Phrase-Level

The statistical results showed that there was a strong main effect of Tune on Minimum F0 ($F(1, 21{,}484.05) = 18{,}613.03$, $p < .001$), with questions showing consistently higher minimum F0

than statements. Tone also showed a significant main effect ($F(2, 42.06) = 254.05$, $p < .001$), indicating substantial differences in minimum F0 realisation across three lexical tones. The main effect of Phrase was particularly strong ($F(2, 21{,}471.98) = 1{,}167.48$, $p < .001$), demonstrating that minimum F0 varies systematically across phrasal positions. Additionally, pre-target tone showed a significant main effect ($F(3, 21{,}479.99) = 445.20$, $p < .001$), suggesting that preceding tonal contexts influence minimum F0 realisation. Several significant interactions between variables were observed. The Tune × Phrase interaction was particularly strong ($F(2, 21{,}471.98) = 1{,}152.15$, $p < .001$), as well as The Tone × Phrase interaction ($F(4, 21{,}471.98) = 403.91$, $p < .001$) and The Phrase × Pre-target tone interaction ($F(6, 21{,}471.98) = 119.95$, $p < .001$). A significant three-way interaction of Tone × Phrase × Pre-target tone was also observed ($F(12, 21{,}471.98) = 22.76$, $p < .001$). See Table 4.9 below for the Minimum F0 result.

Table 4.9 Effects of Tune, Tone, Phrase, and Pre-target on Minimum F0

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	15,106,540.97	15,106,540.97	1	21,484.05	18,613.03	< .001***
Tone	412,384.73	206,192.36	2	42.06	254.05	< .001***
Phrase	1,895,082.24	947,541.12	2	21,471.98	1,167.48	< .001***
Pre-target	1,083,998.10	361,332.70	3	21,479.99	445.20	< .001***
Tune:Tone	10,995.58	5,497.79	2	21,482.41	6.77	.0011**
Tune:Phrase	1,870,194.14	935,097.07	2	21,471.98	1,152.15	< .001***
Tone:Phrase	1,311,266.12	327,816.53	4	21,471.98	403.91	< .001***
Tune:Pre-target	14,315.07	4,771.69	3	21,481.80	5.88	< .001***
Tone:Pre-target	60,365.87	10,060.98	6	21,479.96	12.40	< .001***
Phrase:Pre-target	584,137.99	97,356.33	6	21,471.98	119.95	< .001***
Tune:Tone:Phrase	12,918.19	3,229.55	4	21,471.98	3.98	.0031**
Tune:Tone:Pre-target	1,552.38	258.73	6	21,481.74	0.32	.9275
Tune:Phrase:Pre-target	18,162.62	3,027.10	6	21,471.98	3.73	.001**
Tone:Phrase:Pre-target	221,634.49	18,469.54	12	21,471.98	22.76	< .001***
Tune:Tone:Phrase:Pre-target	5,536.55	461.38	12	21,471.98	0.57	.8691

The post-hoc result showed that when comparing question and statement tunes, questions consistently showed significantly higher minimum F0 than statements across all tones and phrases. This question-statement difference was particularly strong in the post-focus position for all three tones (T1T1: $\beta = 81.30$, $t(21474) = 69.30$, $p < .0001$; T2T2: $\beta = 79.90$, $t(21480) = 68.89$, $p < .0001$; T4T4: $\beta = 77.20$, $t(21474) = 66.42$, $p < .0001$).

As for individual tones, for T1T1 in questions, the post-hoc analyses revealed that minimum F0 was highest in the post-focus position, followed by pre-focus, with focus showing the lowest minimum F0 (pre vs. focus: $\beta = 7.13$, $t(21472) = 6.28$, $p < .0001$; pre vs. post: $\beta = -3.43$, $t(21472) = -3.02$, $p = .01$; focus vs. post: $\beta = -10.56$, $t(21472) = -9.30$, $p < .0001$). In statement contexts with T1T1, minimum F0 decreased progressively from pre-focus to focus to post-focus (pre vs. focus: $\beta = 11.88$, $t(21472) = 9.83$, $p < .0001$; pre vs. post: $\beta = 38.28$, $t(21472) = 31.68$, $p < .0001$; focus vs. post: $\beta = 26.40$, $t(21472) = 21.85$, $p < .0001$).

For T2T2, in questions, minimum F0 increased progressively from pre-focus to focus to post-focus (pre vs. focus: $\beta = -3.41$, $t(21472) = -3.01$, $p = .01$; pre vs. post: $\beta = -10.38$, $t(21472) = -9.14$, $p < .0001$; focus vs. post: $\beta = -6.96$, $t(21472) = -6.14$, $p < .0001$). In statement contexts with T2T2, focus showed the lowest minimum F0, followed by pre-focus, with post-focus showing the highest minimum F0 (pre vs. focus: $\beta = -4.76$, $t(21472) = -4.02$, $p < .001$; pre vs. post: $\beta = 31.40$, $t(21472) = 26.52$, $p < .0001$; focus vs. post: $\beta = 36.15$, $t(21472) = 30.53$, $p < .0001$).

For T4T4, in questions, on-focus showed the lowest minimum F0, followed by post-focus, with pre-focus showing the highest minimum F0 (pre vs. focus: $\beta = -29.89$, $t(21472) = -26.40$, $p < .0001$; pre vs. post: $\beta = -2.85$, $t(21472) = -2.52$, $p = .03$; focus vs. post: $\beta = 27.03$, $t(21472) =$

23.88, $p < .0001$). In statement contexts with T4T4, minimum F0 decreased progressively from pre-focus to focus to post-focus (pre vs. focus: $\beta = -30.45$, $t(21472) = -25.55$, $p < .0001$; pre vs. post: $\beta = 33.68$, $t(21472) = 28.26$, $p < .0001$; focus vs. post: $\beta = 64.13$, $t(21472) = 53.81$, $p < .0001$). See Figure 4.19 below for the effect of tune (left panel) and the effect of phrase (right panel) averaged by pre-target tone. See Table A.1 in Appendix A for the detailed mean minimum F0 across tone, tune, phrase and pre-target tone.

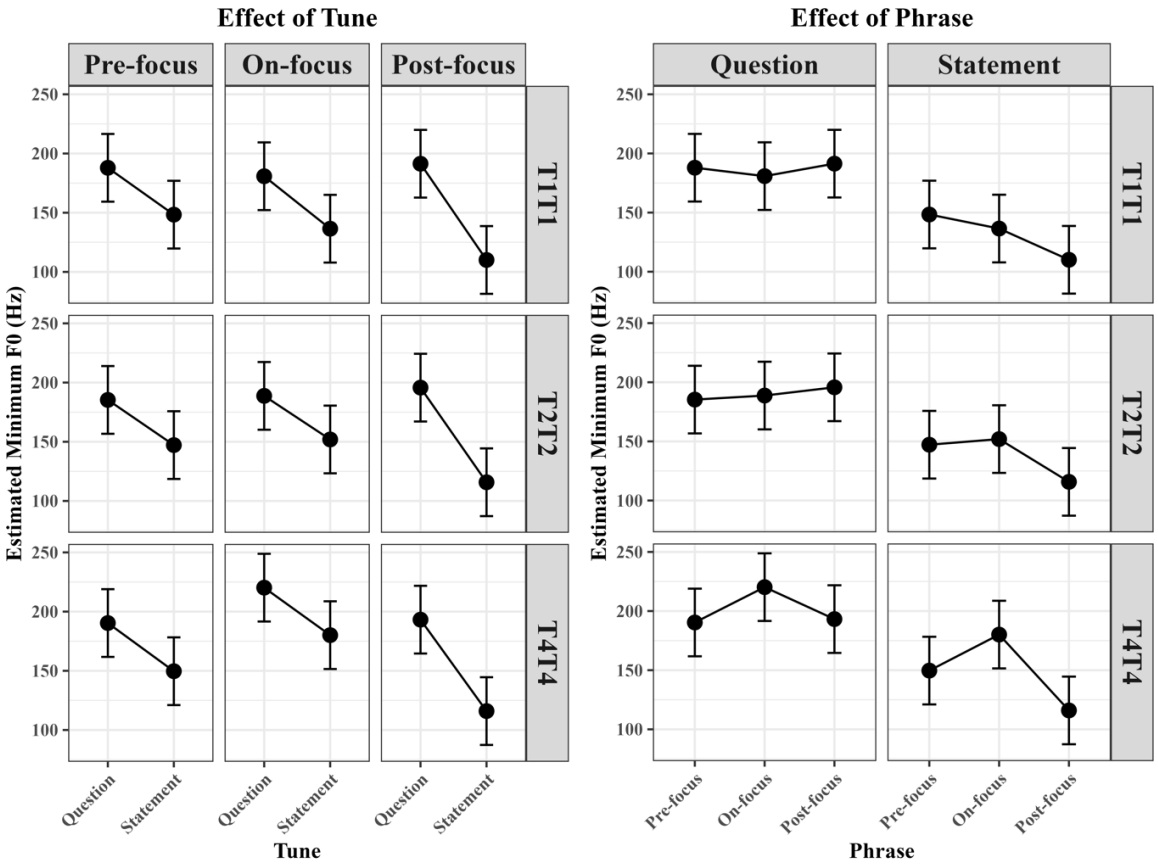


Figure 4.19 Effects of Tune (left panel) and Phrase (right panel) on Minimum F0: Estimated marginal means with 95% confidence intervals

With the addition of the pre-target tone factor (as shown in Figure 4.20 below), the question-statement contrasts show that questions consistently display substantially higher minimum F0

than statements across all tonal combinations, phrasal positions, and pre-tonal contexts. This difference remained the most pronounced in the post-focus position, where the question-statement differences reach their maximum, ranging from approximately 75 Hz to 82 Hz (all $p < .0001$). For example, with pre-target tone T1 and T1T1, the question-statement contrast in post-focus position was 80.70 Hz ($t(21475) = 34.53$, $p < .0001$), while in pre-focus and focus positions the contrasts were 36.30 Hz and 42.90 Hz, respectively. This increase in the question-statement contrast for post-focus position held consistently across all tone types (T1T1, T2T2, T4T4) and pre-target tones (T1, T2, T3, T4), suggesting a fundamental difference in how intonational tunes affect minimum F0 towards the end of utterances, meaning that while pre-target tone does influence minimum F0, the large intonational effect in post-focus position appears relatively resistant to pre-tonal modulation.

The comparisons across phrasal positions showed systematic differences in how minimum F0 values are distributed in questions versus statements. In the question tune, the general pattern showed pre-focus having higher minimum F0 than post-focus position, with significant negative pre-post contrasts across multiple tonal combinations. For instance, with T2T2 and pre-target tone T1, pre-focus had significantly lower minimum F0 than post-focus ($\beta = -30.06$, $t(21472) = -13.22$, $p < .0001$). This pattern was reversed in the statement tune, where pre-focus position typically showed higher minimum F0 than post-focus position, with significant positive pre-post contrasts. For T2T2 with pre-target tone T1 in the statement tune, pre-focus had higher minimum F0 than post-focus ($\beta = 18.59$, $t(21472) = 7.91$, $p < .0001$).

For T1T1, pre-target tone T4 (high-level tone) created distinctly different patterns than pre-target tone T1 (low falling tone). With pre-target tone T4, pre-focus position had significantly higher minimum F0 than focus position in both tunes ($\beta = 17.21$, $t(21472) = 7.58$, $p < .0001$ in

the question tune; $\beta = 17.83$, $t(21472) = 7.38$, $p < .0001$ in the statement tune). With pre-target tone T1, these differences were smaller and non-significant in both tunes. For T4T4, the pre-tonal influence was particularly evident in the focus-post contrast. With pre-target tone T2 (rising tone), focus had significantly higher minimum F0 than post-focus in the question tune ($\beta = 40.53$, $t(21472) = 17.94$, $p < .0001$), while with pre-target tone T1 (low falling tone), focus and post-focus showed similar minimum F0 values ($\beta = -1.32$, $t(21472) = -0.58$, $p = .83$). See Table A.1 in Appendix A for the detailed mean minimum F0 across tone, tune, phrase and pre-target tone.

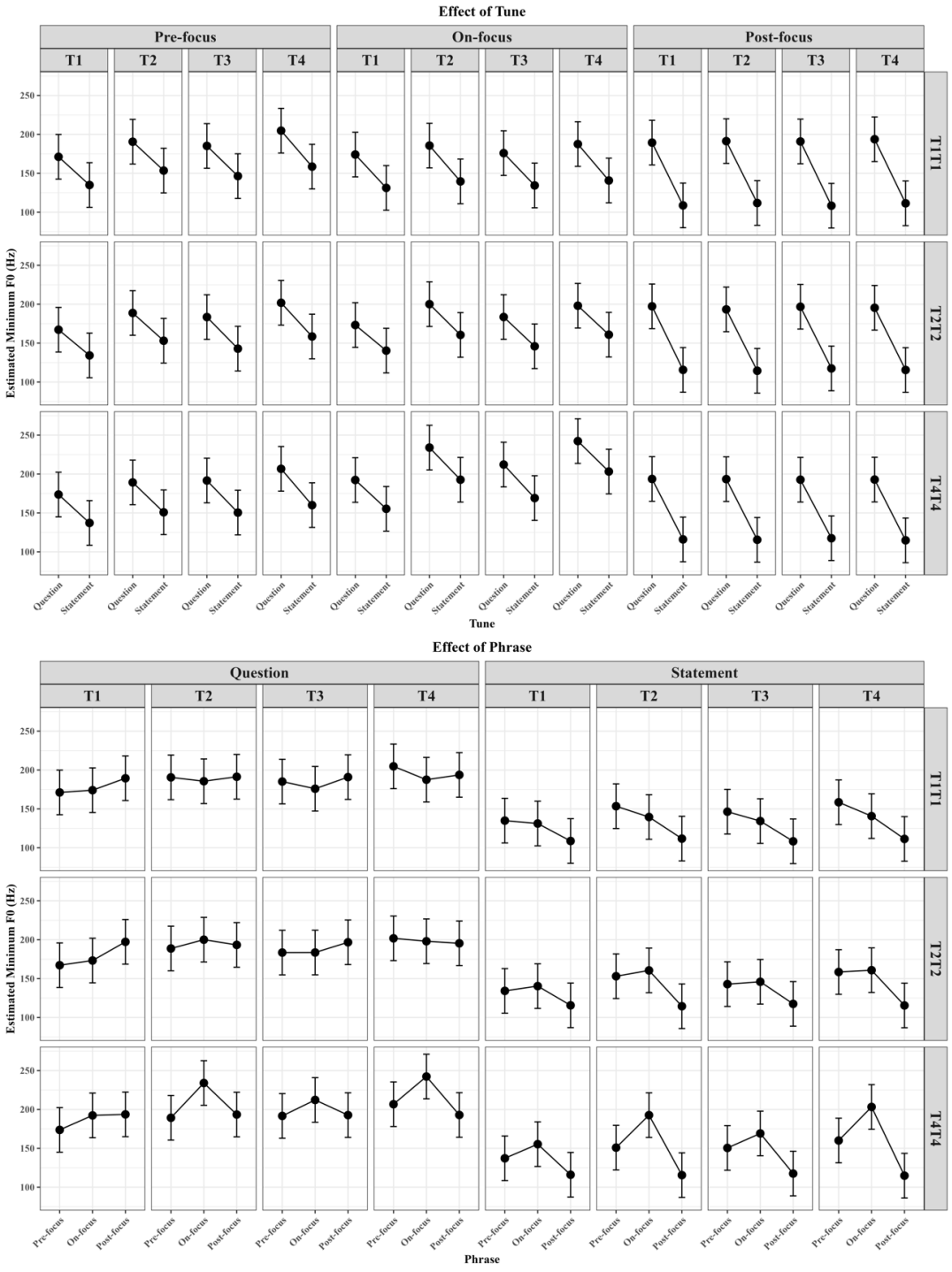


Figure 4.20 Effects of Tune (top panel) and Phrase (bottom panel) on Minimum F0 across pre-target tones: Estimated marginal means with 95% confidence intervals

4.4.7.2 Syllable-Level (On-Focus)

The main effects of Tune (F(1, 14,344.97) = 8,420.04, p < .001), Tone (F(2, 43.98) = 572.58, p < .001), and Syllable (F(1, 14,341.94) = 4,060.23, p < .001) were found significant. The main effect of Pre-target tone was also significant (F(3, 14,343.75) = 222.76, p < .001), as well as interactions, including Tune × Tone (F(2, 14,344.35) = 6.57, p = .001), Tune × Syllable (F(1, 14,341.94) = 289.93, p < .001), and a particularly strong Tone × Syllable interaction (F(2, 14,341.94) = 865.84, p < .001). The three-way interaction of Tune × Tone × Syllable was also significant (F(2, 14,341.94) = 20.54, p < .001). See Table 4.10 for the detailed results.

Table 4.10 Effects of Tune, Tone, Syllable, and Pre-target on Minimum F0

Factor	SS	MS	df (Num)	df (Den)	F	p
Tune	7,717,876.83	7,717,876.83	1	14,344.97	8,420.04	< .001***
Tone	1,049,662.10	524,831.05	2	43.98	572.58	< .001***
Syllable	3,721,635.04	3,721,635.04	1	14,341.94	4,060.23	< .001***
Pre-target	612,544.94	204,181.65	3	14,343.75	222.76	< .001***
Tune:Tone	12,052.13	6,026.06	2	14,344.35	6.57	.0014**
Tune:Syllable	265,752.70	265,752.70	1	14,341.94	289.93	< .001***
Tone:Syllable	1,587,274.25	793,637.12	2	14,341.94	865.84	< .001***
Tune:Pre-target	1,597.40	532.47	3	14,344.19	0.58	.6275
Tone:Pre-target	120,346.64	20,057.77	6	14,343.74	21.88	< .001***
Syllable:Pre-target	580,525.77	193,508.59	3	14,341.94	211.11	< .001***
Tune:Tone:Syllable	37,658.13	18,829.06	2	14,341.94	20.54	< .001***
Tune:Tone:Pre-target	3,047.63	507.94	6	14,344.18	0.55	.7671
Tune:Syllable:Pre-target	6,747.57	2,249.19	3	14,341.94	2.45	.0613
Tone:Syllable:Pre-target	82,458.12	13,743.02	6	14,341.94	14.99	< .001***
Tune:Tone:Syllable:Pre-target	2,841.87	473.64	6	14,341.94	0.52	.7961

The post-hoc analyses comparing the tunes found that questions consistently showed significantly higher minimum F0 than statements across all tone types and syllable positions. For T1T1, this question-statement difference was more pronounced in the second syllable (β = 60.80, t(14342) = 48.76, p < .001) than in the first syllable (β = 36.50, t(14342) = 29.29, p < .001).

For T2T2, the question-statement difference was also larger in the second syllable (β = 56.10, t(14345) = 45.47, p < .001) than in the first syllable (β = 37.30, t(14343) = 30.27, p < .001). For T4T4, the question-statement difference followed the same pattern, being larger in the second syllable (β = 48.50, t(14343) = 39.22, p < .001) than in the first syllable (β = 39.90, t(14343) = 32.26, p < .001).

As for the individual tone comparisons, for T1T1, in question contexts, the second syllable showed significantly lower minimum F0 than the first syllable (β = -14.99, t(14342) = -12.42, p < .001). In contrast, for T1T1 in statement contexts, the second syllable exhibited significantly higher minimum F0 than the first syllable (β = 9.29, t(14342) = 7.23, p < .001). For T2T2, the second syllable showed significantly lower minimum F0 than the first syllable in both question contexts (β = -51.60, t(14342) = -42.78, p < .001) and statement contexts (β = -32.85, t(14342) = -26.11, p < .001), though this syllable difference was more pronounced in questions. Similarly, for T4T4, the second syllable showed significantly lower minimum F0 than the first syllable in both question contexts (β = -55.84, t(14342) = -46.42, p < .001) and statement contexts (β = -47.23, t(14342) = -37.29, p < .001).

When the pre-target tone was added to the analysis, the examination of minimum F0 revealed that pre-target tone has a strong and highly significant influence on pitch realisation, more so than for other F0 measures such as maximum F0. The results showed a main effect of Pre-target tone (F(3, 14,343.75) = 222.76, p < .001), as well as significant interactions with Syllable (F(3, 14,341.94) = 211.11, p < .001) and Tone (F(6, 14,343.74) = 21.88, p < .001). See Table 4.10 for the detailed results.

For T1T1, the pattern is particularly interesting. In the question tune, Syllable 2 consistently shows lower minimum F0 than Syllable 1, with differences ranging from -4.62 Hz (marginally significant, $p = 0.06$) with pre-target tone T2 to -27.50 Hz ($p < .0001$) with pre-target tone T1. However, in the statement tune, this pattern reverses with some pre-target tones: with pre-target tone T2, T3, and T4, Syllable 1 actually shows lower minimum F0 than Syllable 2, with differences ranging from 7.53 Hz to 17.29 Hz (all significant at $p < .01$).

For T2T2, Syllable 2 consistently shows lower minimum F0 than Syllable 1 across all conditions, with substantial differences ranging from -23.57 Hz to -69.60 Hz (all significant at $p < .0001$). This pattern is consistent with the rising contour of T2, where the rise begins from a low pitch point.

For T4T4, the pattern is similar to T2T2 but with even larger differences in some conditions. Syllable 2 consistently shows lower minimum F0 than Syllable 1, with differences ranging from -23.46 Hz to -84.92 Hz (all significant at $p < .0001$). This large difference is somewhat surprising for a high-level tone and may indicate a degree of pitch declination or interaction with intonational factors. See Table A.2 in Appendix A for the detailed mean minimum F0 across tone, tune, syllable and pre-target tone.

4.4.8 F0 Range

4.4.8.1 Phrase-level

There was a significant main effect of Tune ($F(1, 21{,}485.26) = 78.90$, $p < .001$), Tone ($F(2, 42.15) = 17.91$, $p < .001$), and Phrase ($F(2, 21{,}472.07) = 308.46$, $p < .001$) on F0 range. The main effect of Pre-target tone was also significant ($F(3, 21{,}480.85) = 112.01$, $p < .001$). Several significant interactions were also found, including Tune × Tone ($F(2, 21{,}483.49) = 13.68$, p

< .001), Tune × Phrase (F(2, 21,472.07) = 32.90, p < .001), and a particularly strong Tone × Phrase interaction (F(4, 21,472.07) = 338.22, p < .001). The three-way interaction of Tune × Tone × Phrase was also highly significant (F(4, 21,472.07) = 42.42, p < .001). See Table 4.11 below for the detailed results.

Table 4.11 Effects of Tune, Tone, Phrase, and Pre-target on F0 Range

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	249,426.87	249,426.87	1	21,485.26	78.90	< .001***
Tone	113,253.29	56,626.65	2	42.15	17.91	< .001***
Phrase	1,950,156.15	975,078.08	2	21,472.07	308.46	< .001***
Pre-target	1,062,237.61	354,079.20	3	21,480.85	112.01	< .001***
Tune:Tone	86,513.52	43,256.76	2	21,483.49	13.68	< .001***
Tune:Phrase	208,013.59	104,006.80	2	21,472.07	32.90	< .001***
Tone:Phrase	4,276,629.84	1,069,157.46	4	21,472.07	338.22	< .001***
Tune:Pre-target	5,665.18	1,888.39	3	21,482.86	0.60	.6167
Tone:Pre-target	105,201.51	17,533.58	6	21,480.84	5.55	< .001***
Phrase:Pre-target	662,892.67	110,482.11	6	21,472.07	34.95	< .001***
Tune:Tone:Phrase	536,333.70	134,083.43	4	21,472.07	42.42	< .001***
Tune:Tone:Pre-target	13,425.47	2,237.58	6	21,482.79	0.71	.6433
Tune:Phrase:Pre-target	48,803.02	8,133.84	6	21,472.07	2.57	.0171*
Tone:Phrase:Pre-target	217,576.65	18,131.39	12	21,472.07	5.74	< .001***
Tune:Tone:Phrase:Pre-target	61,273.42	5,106.12	12	21,472.07	1.62	.0798

When comparing question and statement tunes, the differences in F0 range varied by tone and phrase position. For T1T1, questions showed significantly smaller F0 range than statements in both pre-focus (β = -6.26, t(21474) = -2.71, p = .01) and focus positions (β = -8.80, t(21474) = -3.80, p < .001), but significantly larger F0 range in post-focus (β = 8.78, t(21474) = 3.80, p < .001). For T2T2, questions showed significantly smaller F0 range than statements in pre-focus (β = -4.87, t(21480) = -2.13, p = .03) and post-focus positions (β = -26.51, t(21480) = -11.58, p < .0001), but significantly larger F0 range in focus (β = 11.97, t(21480) = 5.23, p < .0001). For T4T4, questions showed significantly smaller F0 range than statements in pre-focus (β = -8.46, t(21474) = -3.69, p < .001) and post-focus positions (β = -26.02, t(21474) = -

11.34, $p < .0001$), with no significant difference in focus position ($\beta = -1.26$, $t(21474) = -0.55$, $p = .58$).

Consideration of tonal differences in T1T1 sequences in question contexts revealed that the F0 range was significantly larger in the focus than in the pre-focus ($\beta = -21.62$, $t(21472) = -9.65$, $p < .0001$) and post-focus positions ($\beta = 22.01$, $t(21472) = 9.82$, $p < .0001$), with no significant difference between pre-focus and post-focus ($\beta = 0.39$, $t(21472) = 0.17$, $p = .98$). In statement contexts with T1T1, F0 range was again largest in the focus position (compared to pre-focus: $\beta = -24.15$, $t(21472) = -10.13$, $p < .0001$; compared to post-focus: $\beta = 39.59$, $t(21472) = 16.60$, $p < .0001$), but the post-focus position showed significantly smaller F0 range than the pre-focus position ($\beta = 15.44$, $t(21472) = 6.47$, $p < .0001$).

For T2T2 in question contexts, F0 range increased progressively from pre-focus to focus to post-focus (pre vs. focus: $\beta = -10.58$, $t(21472) = -4.72$, $p < .0001$; pre vs. post: $\beta = -21.25$, $t(21472) = -9.48$, $p < .0001$; focus vs. post: $\beta = -10.67$, $t(21472) = -4.76$, $p < .0001$). In statement contexts with T2T2, F0 range was smallest in the focus position, significantly smaller than both the pre-focus ($\beta = 6.26$, $t(21472) = 2.68$, $p = .02$) and post-focus positions ($\beta = -49.15$, $t(21472) = -21.03$, $p < .0001$), with post-focus showing significantly larger F0 range than pre-focus ($\beta = -42.89$, $t(21472) = -18.35$, $p < .0001$).

For T4T4 in question contexts, F0 range was significantly larger in the post-focus than in both the pre-focus ($\beta = -34.97$, $t(21472) = -15.65$, $p < .0001$) and focus positions ($\beta = -36.97$, $t(21472) = -16.55$, $p < .0001$), with no significant difference between pre-focus and focus ($\beta = 2.01$, $t(21472) = 0.90$, $p = .64$). In statement contexts with T4T4, F0 range increased progressively from focus to pre-focus to post-focus (focus vs. pre-focus: $\beta = 9.21$, $t(21472) = 3.92$, $p < .001$;

pre-focus vs. post-focus: β = -52.52, t(21472) = -22.33, p < .0001; focus vs. post-focus: β = -61.73, t(21472) = -26.24, p < .0001). Figure 4.21 below shows the effect of tune (left panel) and the effect of phrase (right panel) averaged by the pre-target tone.

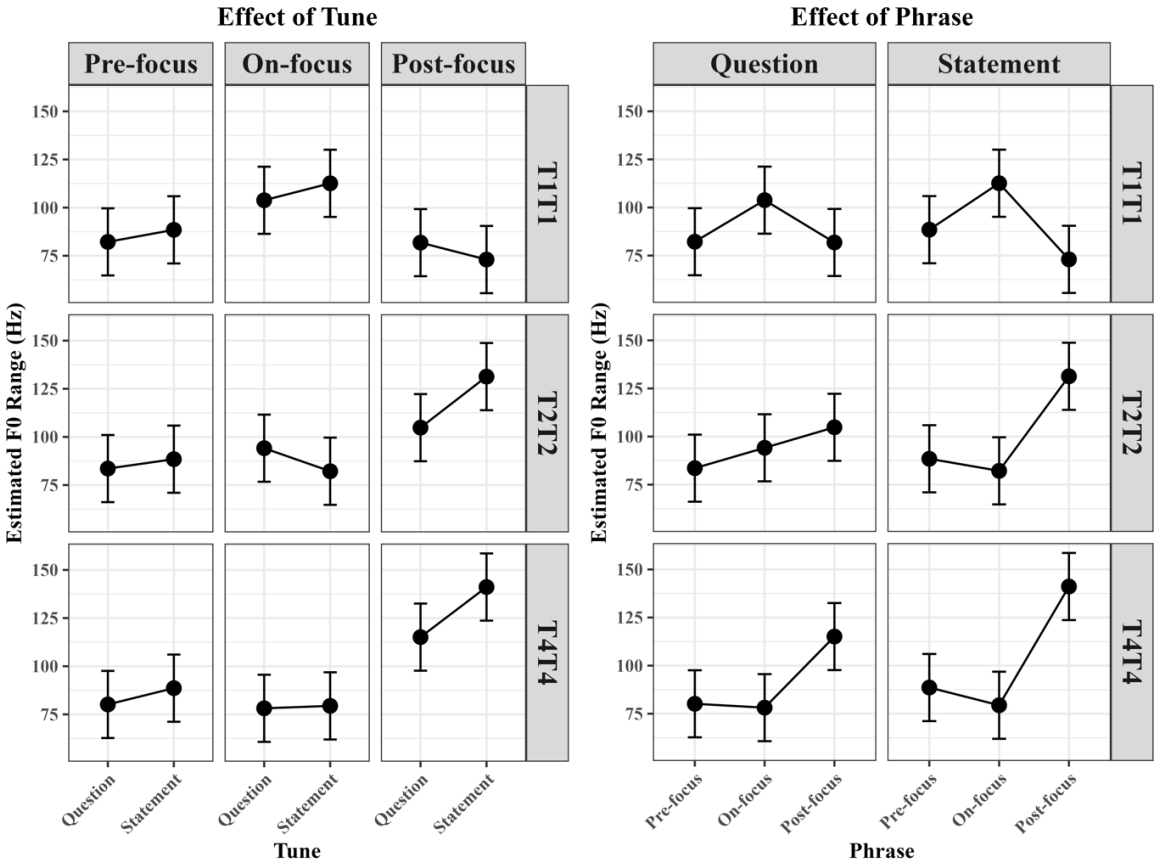


Figure 4.21 Effects of Tune (left) and Phrase (right) on F0 Range: Estimated marginal means with 95% confidence intervals

When the pre-target tone was added as a factor, the post-hoc results showed that the effect of pre-target tone is not uniform on the F0 range on the preceding tones. The interaction between pre-target tone and phrase was particularly strong, F(6, 21,472.07) = 34.95, p < .001, indicating that the effect of pre-target tone differs markedly across pre-focus, on-focus, and post-focus positions. Similarly, Tone × pre-target tone was significant, F(6, 21,480.84) = 5.55, p < .001,

confirming that the effect of pre-target tone varies depending on the identity of the following tone.

In pre-focus positions, speakers generally produced a smaller F0 range compared to focus, but the extent of this reduction depended on the preceding tone. For instance, when the pre-target tone was a low falling tone (T1) and the following tone was the rising-falling T1T1, the F0 range in pre-focus was significantly narrower than in focus, β = -18.76, t(21,472) = -4.20, p < .001. In contrast, when the following tone was T4T4 (high-level), the difference between pre- and focus was more moderate, β = -11.86, t(21,472) = -2.66, p = .02. In some cases, such as T4T4 with T3 as pre-target tone, no reliable pre-focus compression was observed, p = .61.

In on-focus phrases, pre-target tone again modulated the pitch range, often enhancing or attenuating the degree of pitch excursion. For example, when the pre-target tone was the rising tone T2 and the upcoming tone was T1T1, F0 range in focus was significantly smaller than in post-focus, β = 15.57, t(21,472) = 3.47, p = .002, suggesting a subtle carryover of pitch height or pitch register. A similar pattern was found with other tone combinations, such as T2 pre-target tone and T2T2, where F0 range dropped sharply from focus to post-focus, β = -23.52, t(21,472) = -5.23, p < .001.

In post-focus phrases, pre-target tone effects were most evident in the extent of post-focus compression. For instance, when the pre-target tone was T4 (high-level) and the upcoming tone was T4T4, the post-focus F0 range was substantially reduced relative to both pre- and on-focus positions, β = -60.40, t(21,472) = -13.57, p < .001. In contrast, when the pre-target tone was T1 (low falling), the degree of post-focus compression was generally milder, especially for T1T1 tones, β = -7.55, t(21,472) = -1.69, p = .21. See Figure 4.22 for the effect of tune and phrase

across different pre-target tones and Table A.1 in Appendix A for the detailed mean F0 range across tone, tune, phrase and pre-target tone.

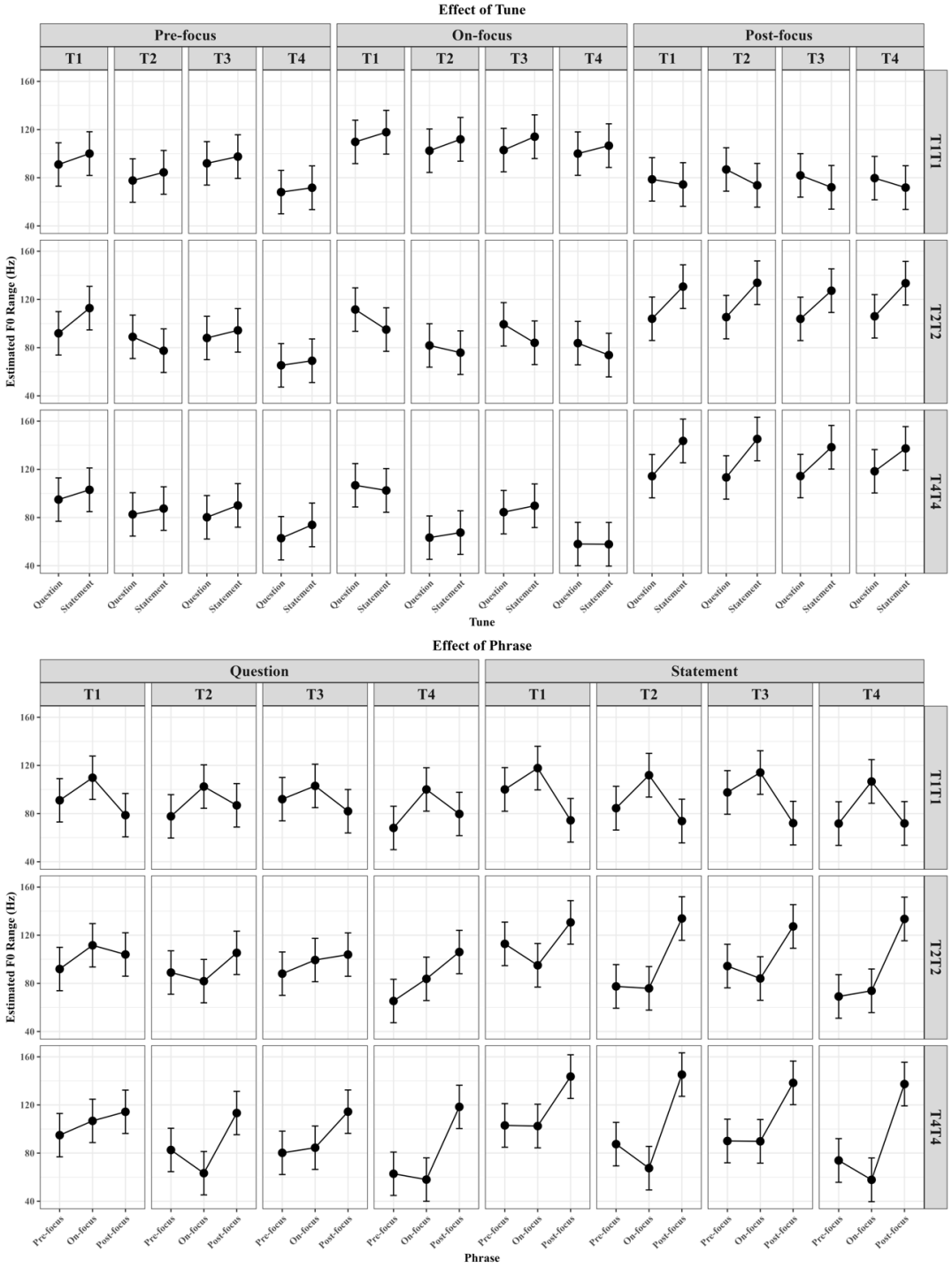


Figure 4.22 Effects of Tune (top panel) and Phrase (bottom panel) on F0 Range across pre-target tones: Estimated marginal means with 95% confidence intervals

4.4.8.2 Syllable-level (On-Focus)

The results for F0 range revealed significant effects for all main factors at the syllable level: Tune ($F(1, 14{,}343.44) = 10.73$, $p = .001$), Tone ($F(2, 44.02) = 113.80$, $p < .001$), Syllable ($F(1, 14{,}341.96) = 2{,}056.65$, $p < .001$), and pre-target tone ($F(3, 14{,}342.79) = 148.55$, $p < .001$). Several interactions are also highly significant, including Tune × Tone ($F(2, 14{,}343.07) = 35.03$, $p < .001$), Tune × Syllable ($F(1, 14{,}341.96) = 197.63$, $p < .001$), Tone × Syllable ($F(2, 14{,}341.96) = 355.08$, $p < .001$), Tone × pre-target tone ($F(6, 14{,}342.79) = 9.43$, $p < .001$), and Syllable × pre-target tone ($F(3, 14{,}341.96) = 118.08$, $p < .001$). See Table 4.12.

Table 4.12 Effects of Tune, Tone, Syllable, and Pre-target on F0 Range

Factor	SS	MS	df (Num)	df (Den)	*F*	*p*
Tune	12,288.62	12,288.62	1	14,343.44	10.73	.0011**
Tone	260,547.07	130,273.54	2	44.02	113.80	< .001***
Syllable	2,354,339.64	2,354,339.64	1	14,341.96	2,056.65	< .001***
Pre-target	510,173.54	170,057.85	3	14,342.79	148.55	< .001***
Tune:Tone	80,206.53	40,103.26	2	14,343.07	35.03	< .001***
Tune:Syllable	226,237.98	226,237.98	1	14,341.96	197.63	< .001***
Tone:Syllable	812,960.94	406,480.47	2	14,341.96	355.08	< .001***
Tune:Pre-target	3,309.56	1,103.19	3	14,343.00	0.96	.4088
Tone:Pre-target	64,789.79	10,798.30	6	14,342.79	9.43	< .001***
Syllable:Pre-target	405,510.64	135,170.21	3	14,341.96	118.08	< .001***
Tune:Tone:Syllable	35,025.62	17,512.81	2	14,341.96	15.30	< .001***
Tune:Tone:Pre-target	13,933.53	2,322.25	6	14,343.00	2.03	.0583
Tune:Syllable:Pre-target	3,806.73	1,268.91	3	14,341.96	1.11	.3442
Tone:Syllable:Pre-target	83,959.32	13,993.22	6	14,341.96	12.22	< .001***
Tune:Tone:Syllable:Pre-target	1,386.90	231.15	6	14,341.96	0.20	.9763

When averaged by the pre-target tone, the post-hoc results showed that when comparing question and statement tunes, the differences in F0 range varied by tone and syllable position. For T1T1, questions showed significantly larger F0 range than statements in the first syllable ($\beta = 4.57$, $t(14342) = 3.28$, $p = .001$), but significantly smaller F0 range in the second syllable

(β = -19.96, t(14342) = -14.33, p < .001). For T2T2, questions showed significantly larger F0 range than statements in the first syllable (β = 10.52, t(14343) = 7.63, p < .001), but slightly smaller F0 range in the second syllable (β = -2.70, t(14343) = -1.96, p = .05). For T4T4, questions showed significantly larger F0 range than statements in the first syllable (β = 3.16, t(14342) = 2.29, p = .02), but significantly smaller F0 range in the second syllable (β = -6.72, t(14342) = -4.87, p < .001). Thus, across all three tones, questions consistently showed larger F0 range than statements in the first syllable, but smaller F0 range in the second syllable.

For T1T1 with the question intonation, the first syllable showed significantly larger F0 range than the second syllable (β = 18.70, t(14342) = 13.87, p < .001). In contrast, for T1T1 in statement contexts, the second syllable exhibited significantly larger F0 range than the first syllable (β = -5.83, t(14342) = -4.07, p < .001). For T2T2, the first syllable showed significantly larger F0 range than the second syllable in both question contexts (β = 33.72, t(14342) = 25.01, p < .001) and statement contexts (β = 20.50, t(14342) = 14.58, p < .001), though this syllable difference was more pronounced in questions. Similarly, for T4T4, the first syllable showed significantly larger F0 range than the second syllable in both question contexts (β = 48.24, t(14342) = 35.89, p < .001) and statement contexts (β = 38.36, t(14342) = 27.10, p < .001), with this syllable difference again being more obvious in questions.

When the pre-target tone was added as a factor, the post-hoc tests showed that for T1T1 in the question tune, Syllable 1 consistently shows a larger F0 range than Syllable 2 across all pre-target tone conditions, with differences ranging from 12.13 Hz with pre-target tone T2 to 28.98 Hz with pre-target tone T1 (all significant at p < .0001). However, in the statement tune, this pattern is more variable: with pre-target tone T1, the difference is small and non-significant (2.87 Hz, p = 0.31), while with pre-target tones T2, T3, and T4, the pattern actually reverses,

with Syllable 2 showing a larger F0 range than Syllable 1 (differences of -8.24 Hz, -6.14 Hz, and -11.83 Hz, respectively, all significant at $p < .05$). This tune-dependent reversal suggests that tune significantly alters how tonal contours are realised, particularly for T1T1 sequences. For T2T2, Syllable 1 consistently shows a larger F0 range than Syllable 2 across all conditions, though the magnitude varies considerably. In the question tune, the differences range from 24.31 Hz with pre-target tone T2 to 48.59 Hz with pre-target tone T1. In the statement tune, the differences are generally smaller, ranging from 11.97 Hz with pre-target tone T4 to 32.27 Hz with pre-target tone T1. For T4T4, the pattern is similar to T2T2, with Syllable 1 consistently showing a larger F0 range than Syllable 2. The differences are particularly large under some conditions, reaching 73.95 Hz with pre-target tone T1 in the question tune and 61.35 Hz with pre-target tone T1 in the statement tune. See Table A.2 in Appendix A for the detailed mean F0 range across tone, tune, syllable and pre-target tone.

4.4.9 Maximum F0 and Minimum F0 Alignment

The precise location of maximum and minimum F0 within each utterance was first identified to analyse the distribution patterns of F0 extrema. These locations were then subsequently categorised by phrasal region (pre-focus, on-focus, or post-focus). For each combination of Tone (T1T1, T2T2, or T4T4) and Tune, occurrences of F0 maxima and minima in each phrasal region were tallied. These raw counts were then converted to percentages by dividing by the total number of occurrences within each Tone-Tune combination and multiplying by 100. The maximum F0 distribution by phrase is presented in Figure 4.23 and with pre-target tone in Figure 4.24, and minimum F0 distribution by phrase is presented in Figure 4.25 and with pre-target tone in Figure 4.26.

At the whole-utterance level (Figure 4.23), the alignment between maximum F0 and focus position is tone-dependent. For T1T1 and T4T4, the pitch peak most frequently occurred in the on-focus position, consistent with a pitch accent aligning with the focused element, which is a tendency that is particularly noticeable in statements. T2T2, however, displays the opposite pattern: on-focus is the least likely site of maximum F0 in both questions (22.06%) and statements (23.40%), with the peak instead favouring the post-focus or peripheral regions.

Regarding T1T1, maximum F0 most commonly occurs in the on-focus position, with a 41.06% likelihood in questions and an even higher 49.46% in statements. The pre-focus position also shows a greater probability, 31.22% in questions and 39.21% in statements, while the post-focus region has relatively lower occurrence rates.

For T2T2, questions depart from this pattern, with maximum F0 most frequently occurring in the post-focus region (45.79%), followed by pre-focus (32.14%) and the lowest proportion on-focus (22.06%). In statements, the distribution becomes more balanced between pre-focus (38.60%) and post-focus (38.00%), with on-focus remaining the least likely site at 23.40%.

T4T4 displays the strongest alignment between maximum F0 and the focused position, with the on-focus region most frequently hosting the pitch peak in both questions (51.07%) and statements (58.27%). In questions, pre-focus and post-focus regions are nearly equal in distribution (24.31% and 24.63%, respectively), while in statements, post-focus drops further (16.10%) compared to pre-focus (25.63%).

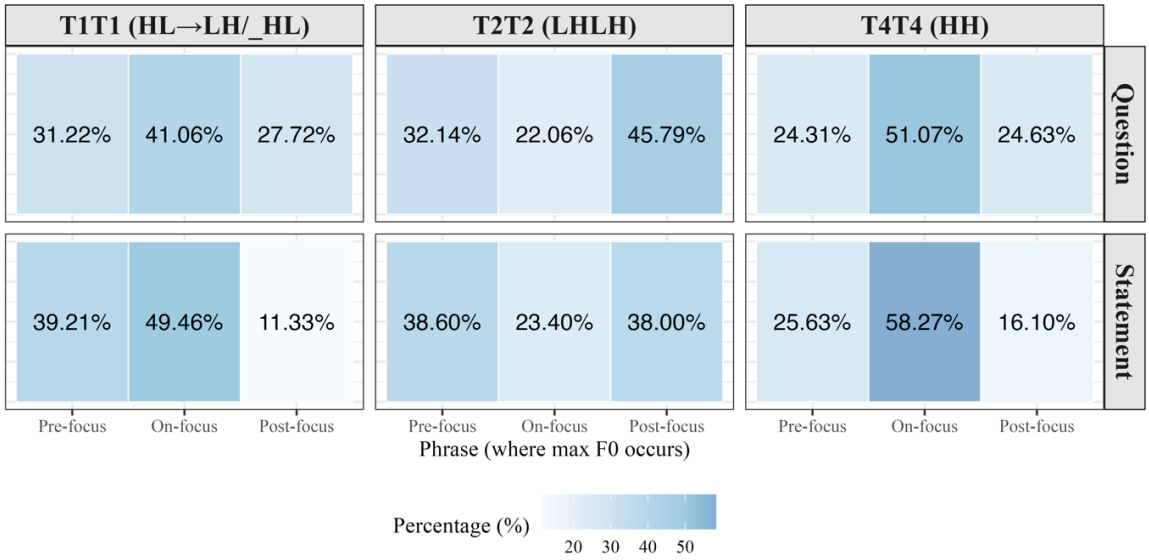


Figure 4.23 Maximum F0 distribution by phrase in questions and statements

When considering pre-target tones, as shown in Figure 4.24 below, the lexical tone preceding the focus, in T1T1, T1 and T2 pre-targets are associated with higher probabilities of on-focus maximum F0 in questions (48.58% and 44.90%, respectively), while T3 pre-targets show a stronger preference for pre-focus peaks (43.45%). Statements within this pattern show consistently high on-focus probabilities across all pre-target tones (ranging from 43.82% to 52.03%), with post-focus regions showing the lowest percentages (from 9.59% to 13.31%).

As for T2T2, in questions, post-focus maxima are dominant across all pre-target tones, especially with T4 (51.12%). T3 shows an equal likelihood of maximum F0 occurring in both pre-focus and post-focus regions (42.19% each). In statements, the distribution is variable: T2 pre-targets tend to favour post-focus peaks (42.36%), while T3 pre-targets are more associated with pre-focus maxima (44.48%).

T4T4 demonstrates the most stable and consistent distribution, with strong on-focus maximum F0 alignment across all pre-targets in both questions and statements. This is especially evident in statements with a T4 pre-target, where the on-focus peak reaches 63.08%. Across all patterns, post-focus maximum F0 is generally less common in statements than in questions, suggesting that questions are more likely to maintain high pitch later in the utterance.

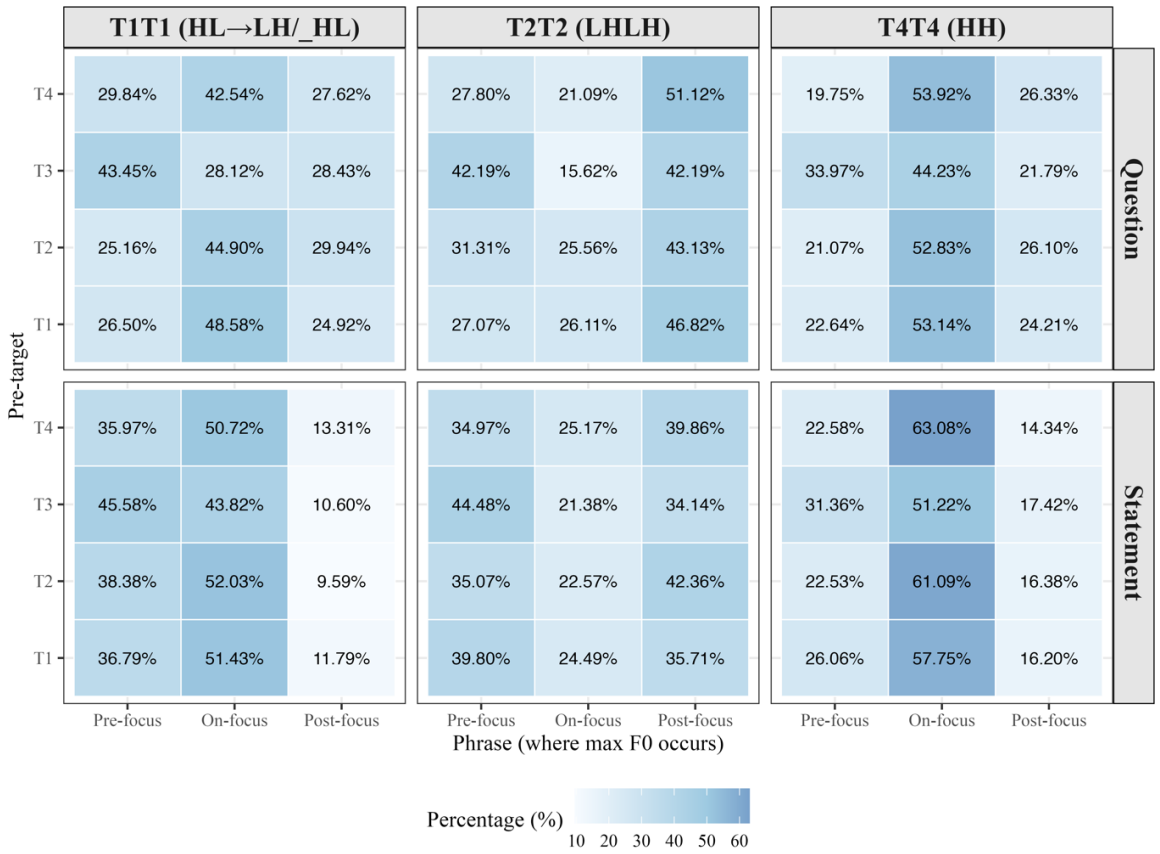


Figure 4.24 Maximum F0 distribution by phrase and pre-target tone in questions and statements

The distribution of minimum F0 in questions (as shown in Figure 4.25, top panel) shows distinct patterns across intonation types. At the whole-utterance level, within T1T1, minimum F0 is most likely to occur in the on-focus position (45.04%), followed by pre-focus (29.47%)

and post-focus (25.50%). In T2T2, the distribution is relatively balanced, with pre-focus slightly more likely (35.63%), followed by on-focus (32.38%) and post-focus (31.98%). T4T4 presents a strikingly different trend: minimum F0 is most commonly located on-focus (51.46%) and post-focus (44.67%), with very little occurring in pre-focus (3.87%).

For statements (bottom panel), T1T1, T2T2, and T4T4 all show a strong concentration of minimum F0 in the post-focus position, 87.05%, 82.47%, and 85.39%, respectively. This points to a strong post-focal pitch lowering across statements, regardless of tone pattern. However, the on-focus position shows extremely low probabilities for minimum F0: 7.01% in T1T1, 5.35% in T2T2, and just 0.70% in T4T4, with the HH tone sequence showing the most extreme avoidance of placing pitch minima on focused elements.

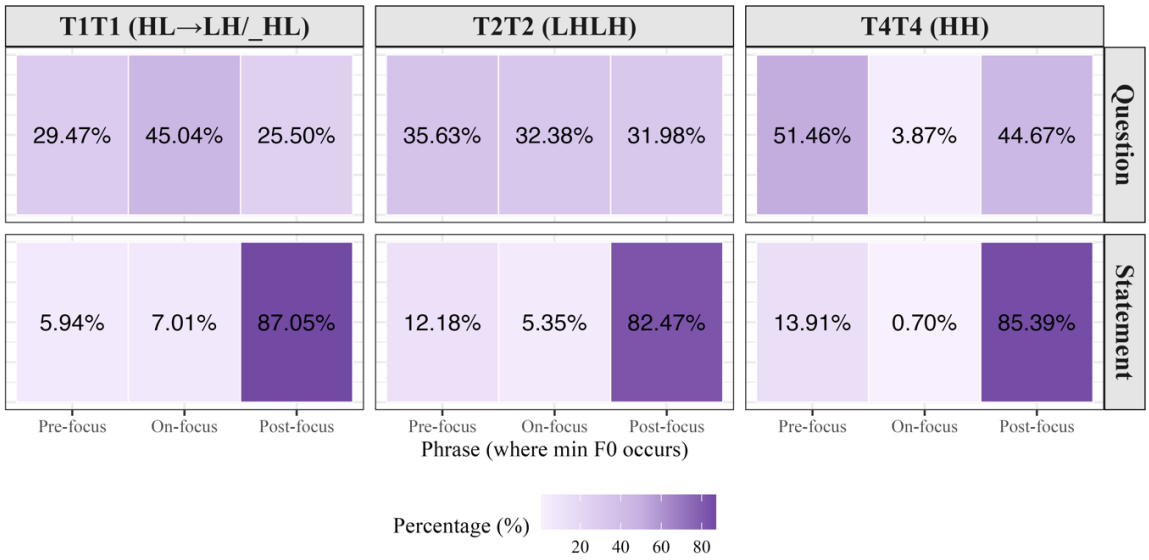


Figure 4.25 Minimum F0 distribution by phrase in questions and statements

Figure 4.26 below provides a detailed view of how minimum F0 distribution is influenced by pre-target tone. While the earlier figure showed general trends across entire utterances, this

breakdown reveals how the lexical tone preceding the focused element affects where pitch valleys are realised within the whole sentence. In questions, the distribution of minimum F0 shows greater variability and is more sensitive to the identity of the pre-target tone. For T1T1, the on-focus region is the most frequent locus of the minimum F0, particularly following T3 (51.12%) and T4 (46.98%) pre-targets, whereas T1 and T2 pre-targets show higher pre-focus probabilities. T2T2 questions are more variable still, with minimum F0 occurring across all positions depending on pre-target tone. Specifically, T1 and T2 show high pre-focus probabilities (44.59% and 45.05%), while T4 shows a post-focus minimum dominance (46.01%), reflecting tonal influence on pitch contour realisation. T4T4, on the other hand, has a clear preference for minimum F0 in the pre-focus region across all tones, most with T1 (72.64%) and T2 (56.60%). The on-focus position is almost entirely avoided in these cases, indicating that pitch valleys are rarely aligned with focal elements in high-level contour questions.

In statements, the patterns are still consistent across all intonation types: minimum F0 almost always occurs in the post-focus region. For the T1T1 pattern, post-focus percentages range from 80.71% with a T1 pre-target to 89.40% with T3. Similarly, in T2T2, post-focus minima are dominant across all tones, rising as high as 90.91% with a T4 pre-target. T4T4 exhibits the most stable distribution of all, with post-focus minimum F0 placement exceeding 76% for all pre-targets, and reaching 91.40% with T4. On-focus and pre-focus occurrences in statements are minimal across all conditions, strongly supporting the presence of post-focal pitch lowering as a universal feature of statement intonation, regardless of tonal context.

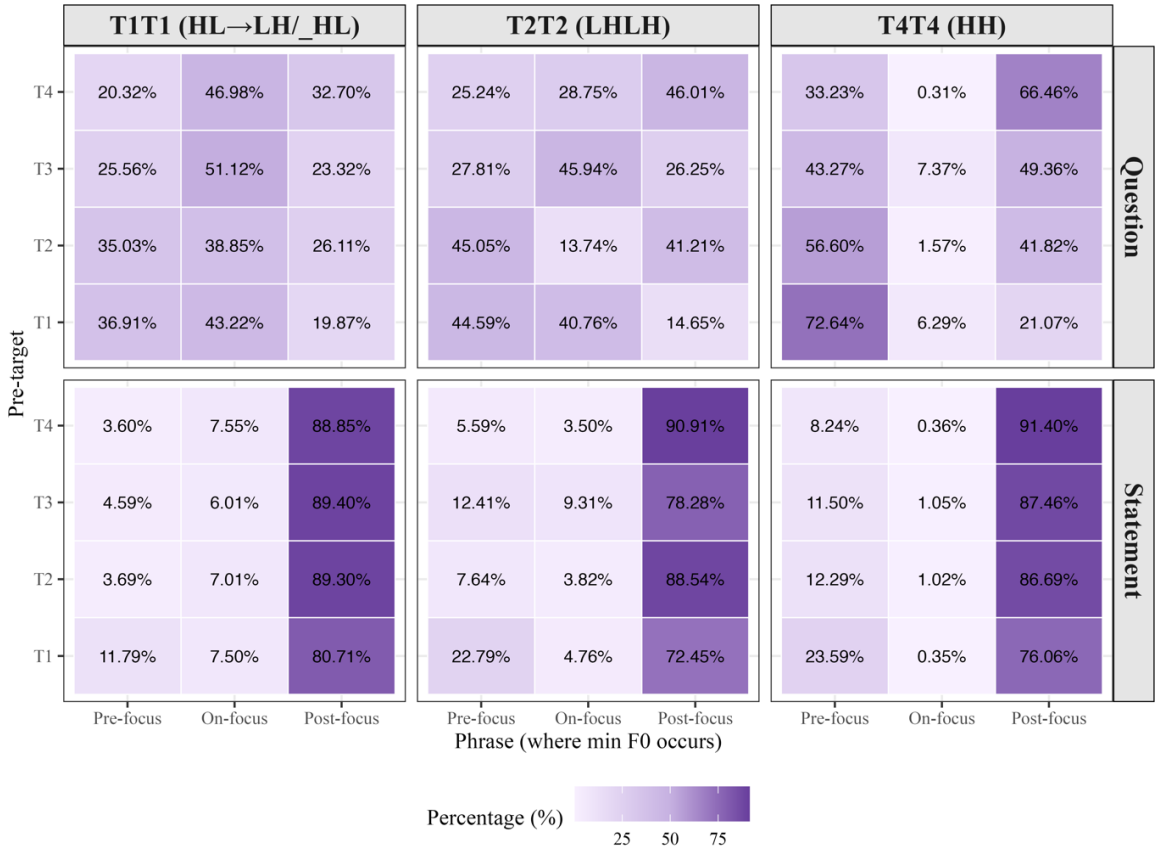


Figure 4.26 Minimum F0 distribution by phrase and pre-target tone in questions and statements

4.5 Discussion

4.5.1 Effects of Pre-Target Tones

The analyses of F0 mean, maximum, minimum, and range collectively suggest that pre-target tones have a clear carry-over effect on the realisation of subsequent tones. Rather than altering the tonal shape or category of the following tone (as shown in Figure 4.4), pre-target tone appears to modulate the overall pitch register, influencing how high or low the tone is produced while preserving its lexical contour. For instance, when the preceding tone was a high-level (T4) or high-falling tone (T3), the subsequent phrase often exhibited elevated F0, including

higher F0 maxima, minima, and F0 means. Meanwhile, low-falling (T1) or rising (T2) pre-target tones were associated with a lower overall pitch level.

This pattern was consistent across tones and phrase positions, with the expected tonal trajectories (e.g., rise-fall for T1T1 or rising for T2T2) remaining intact. Even where F0 range varied as a function of pre-target tones, for example, with compressed or expanded pitch excursions, these changes reflected modulation in pitch span rather than structural alteration of the tonal contour. Thus, the influence of pre-target tones is best characterised as phonetic carryover, whereby tonal height and pitch register are affected by preceding tones, but the phonological identity and shape of the tone remain stable.

4.5.2 Parallel Encoding

4.5.2.1 Prosodic Focus Marking

Prosodic focus is marked primarily by increased intensity and reduced duration, both of which contribute to the perceived salience of the focal element. At both the phrasal and syllabic levels of the on-focus phrase, focal positions consistently showed greater mean intensity, indicating that speakers adjust loudness to mark focus. Since each prosodic region contains an equal number of syllables, the shorter duration of the on-focus region can be interpreted as a locally faster speech rate rather than segmental reduction, a form of temporal compression associated with prominence. Speech rate and intensity therefore emerge as the primary and most consistent phonetic correlates of focus, applying across all three tone combinations irrespective of their pitch configurations.

The role of F0 in focus marking, by contrast, is tone-dependent. For T1T1 and T4T4, maximum F0 aligned most frequently with the on-focus position in both statements and questions (Figure

4.23). In statements, the on-focus region hosted the F0 peak in 49.46% of T1T1 tokens and 58.27% of T4T4 tokens, and in questions it remained the modal location at 41.06% and 51.07%, respectively, suggesting a local high pitch, plausibly an H* pitch accent, aligning with the focused element. The alignment is especially revealing in questions, where two additional F0-raising mechanisms are present that are absent in statements: the globally raised interrogative register and the final high boundary tone (§4.5.2.3).

For T2T2, however, the focused word does not host the utterance-level maximum. In both statements and questions, the on-focus region was the least likely site of maximum F0 (23.40% and 22.06%, respectively), with the peak instead falling most often in the post-focus region. Within these utterances, the maximum F0 of the focused word, around 283 Hz in questions and 237 Hz in statements, remains lower than that of the following post-focus word (around 301 Hz and 250 Hz), so the maximum is located after the focus rather than on it. This tendency is stronger in questions, where the boundary tone raises the utterance-final region further (post-focus maximum in 45.79% of tokens). The focal accent is therefore present but does not surface as the utterance peak.

This does not mean that focus is unmarked in T2T2, but that the lexical tone takes the available F0, leaving the accent unrealised in pitch. For T1T1 and T4T4, the focused word reaches the highest F0 in the utterance, so the H* surfaces as the peak; for T2T2, the focused word does not reach as high as the following post-focus word, so the accent, though present, is not realised in F0, and focus is carried by intensity and speech rate instead. The H* pitch accent is thus best understood as a consistent component of the focus tune, present across all three tones but visible in F0 only where the lexical tone allows, with intensity and rate marking focus reliably throughout.

This pattern stands in contrast to previous work on Mandarin focus. Much of the literature reports focus-related F0 expansion and durational lengthening (e.g. Y. Xu, 1999), and, most directly relevant, Duan and Jia (2014) found that F0 was the primary focus cue for GuanM and that post-focus compression was present. In contrast, the present results for GuanM show speech rate and intensity as the primary cues, with F0 playing a secondary and tone-dependent role. As discussed in previous chapters, the four lexical tones in GuanM (except T4, the high-level tone) are significantly shorter in duration compared to SC (Liu et al., 2020), which may reflect a language-specific tendency towards faster speech delivery. This tendency could underlie the observed shortened duration as a cue to sentential prominence in GuanM.

4.5.2.2 Declination in Statements

No evidence of downstep was found in the present study, as the falling and rising tone combinations used here do not involve any phonological triggers for tonal lowering. Any systematic downtrend in statements must therefore be interpreted as a register-level effect rather than as the cumulative result of local tonal lowering.

A consistent, gradual lowering of F0 is observed across declarative utterances (Figure 4.10). Regardless of the tones in the on-focus region or the nature of the pre-target tones, statements exhibit negative F0 slopes across the utterance, reflecting a low intonational register. The consistency of this effect across tonal contexts supports a register-level interpretation: lowering driven by local tonal interactions would be expected to vary with tonal composition, while a declining register lowers all tones alike. Two further observations support this interpretation. First, Figure 4.9 shows a positive F0 slope at the onset of declaratives against a negative onset slope in interrogatives, indicating a pitch reset at the start of statement utterances; this reset can

be understood as the speaker setting the reference line high, establishing the ceiling from which declination proceeds. Second, F0 valleys in statements occur consistently in post-focus positions, with the lowest pitch values reached after the focal element rather than on any particular tone.

The comparison between sentence types provides the strongest evidence. Questions and statements contain identical lexical tone sequences and differ only in tune, yet the downtrend appears only in statements. This shows that the lowering is not an automatic phonetic tendency, nor a consequence of the tones themselves, but a property of the statement tune specifically. This finding supports Ladd's (1984) view that declination is not inevitable but emerges from intonational structure: here, from a low, declining register present in statements and absent in questions.

Finally, this register-level lowering is not readily captured by the local, target-based representations of AM Theory, which specify pitch through the scaling of individual H and L tones. Because the tonal sequences provide no downstep triggers, the downtrend cannot be derived from stepwise tonal lowering, as shown in (7), and is better analysed as a property of the statement register.

4.5.2.3 Question Intonation

In addition to the prosodic marking of focus, the intonational contour associated with questions exhibits a globally higher F0 register from the onset of the utterance, supporting the view that interrogative tunes are realised with an overall pitch elevation. It is consistent across tone types and phrasal positions, with questions showing higher mean F0, maximum F0, and minimum F0 compared to statements. The difference between questions and statements was most marked

in post-focus positions, suggesting that both the utterance-final boundary tone and the global register setting contribute to the interrogative tune. For clearer visualisation, Figure 4.27 below presents the F0 contours averaged by different pre-target tones to present the overall pitch trend across both intonational types.

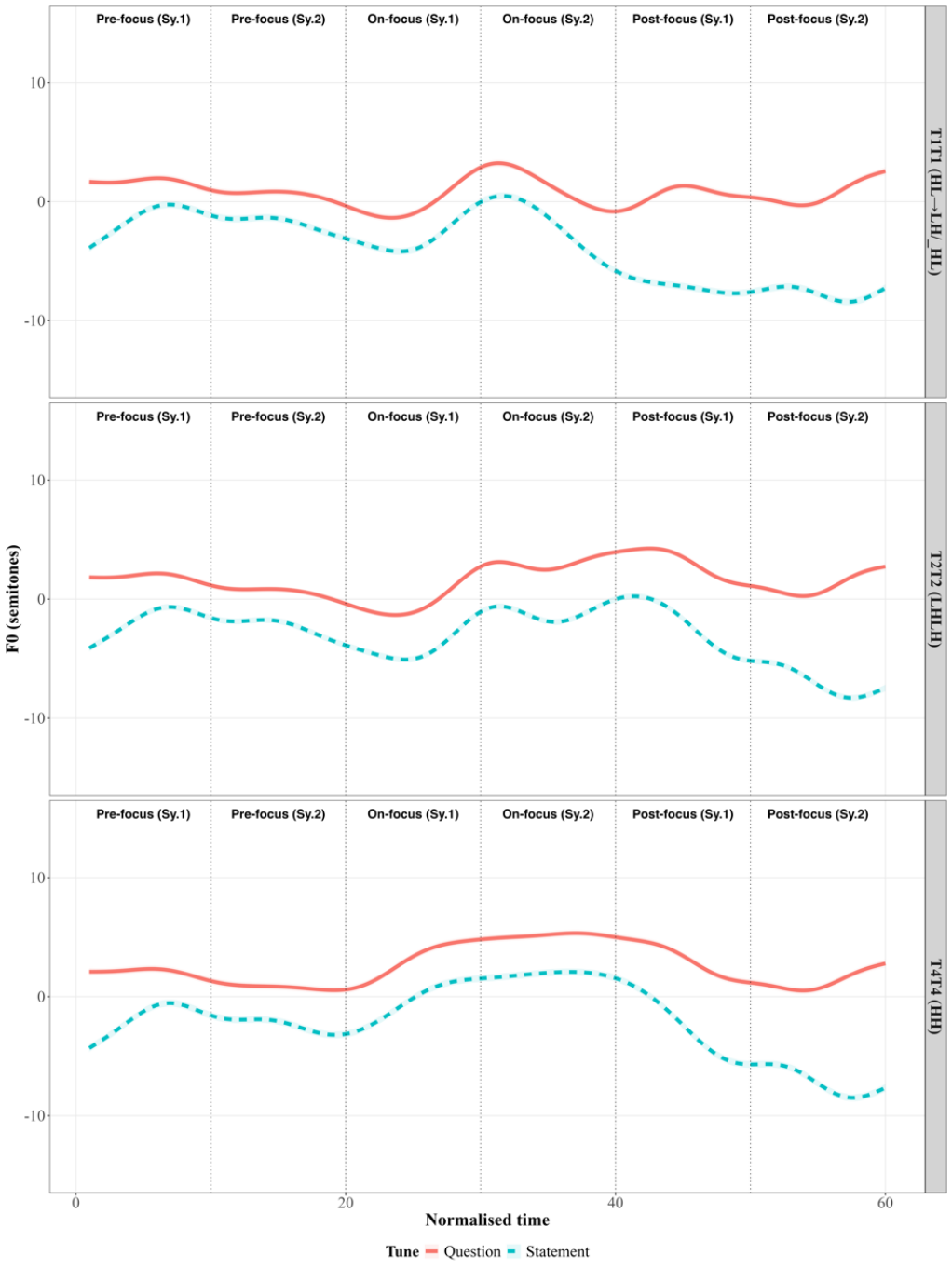


Figure 4.27 Time-normalised smoothed F0 contours of sentences across three tones, fitted using GAMMs with 95% confidence intervals

The observed question contour aligns with findings from Production Study 1, which proposed the presence of a high boundary tone in interrogatives, as well as the globally raised pitch register. The high boundary tone appears to be phonetically realised in a tone-sensitive manner on the final syllable: it does not always produce an overt rising contour. Instead, it tends to enhance the rising trajectory of inherently rising tones and suppress the falling trajectory of falling tones, thereby preventing final syllables from reaching the lower pitch levels typically observed in statements. Crucially, the disyllabic data in longer utterances provide stronger evidence for the edge-aligned nature of the high boundary tone. The F0 difference between questions and statements remained roughly consistent throughout the utterance until the post-focus region. However, in the final prosodic word, comprising two syllables, both syllables showed a marked increase in the question-statement difference. If the boundary tone were a floating influence exerting diffuse upward pressure, the raising effect would be distributed more uniformly across the entire utterance. Instead, the concentration of the effect on the final prosodic word supports the analysis of H% as a fixed target aligned to the IP boundary, with its influence extending across the edge-adjacent domain rather than the phrase as a whole.

The phonetic realisation of the high boundary tone remains tone-dependent, facilitating greater rise in rising tones and mitigating fall in falling tones. Combined with the global pitch elevation, this configuration not only reinforces the interrogative function prosodically, but also contributes to stabilising pitch variation across the entire utterance. In doing so, speakers are able to maintain tonal consistency for lexical tones, ensuring that lexical contrasts remain intact while simultaneously conveying sentence-level meaning through intonation.

4.6 Conclusion

This chapter has examined how prosodic focus and intonational meaning are phonetically realised in longer utterances in GuanM, with particular emphasis on the interaction between lexical tone and sentence-level intonation. The findings demonstrate that focus is primarily marked by faster speech rate and increased intensity, rather than by F0-based acoustic features. This suggests that prominence is achieved not through temporal expansion, as is often reported in other languages, but through prosodic compression and energy enhancement.

One possible explanation for this pattern is that F0 is already functionally occupied by both lexical tone and intonation, leaving focus, another layer of pragmatic meaning, to be realised by alternative phonetic cues. At the same time, the question intonational tune is conveyed through a global high register and a final high boundary tone, whereas statements are characterised by a global low register and declination, with H* pitch accent aligned with focus.

Based on these observations, the intonational tunes for statements and questions in GuanM are summarised as follows:

Statement tune:

Lexical tone sequences + H* pitch accent on focus phrases + global low register with declination

Question tune:

Lexical tone sequences + H* pitch accent on focus phrases (which may be less perceptible) + global register raising + final H% boundary tone

Chapter 5 Perception of Question Intonation in Guanzhong Mandarin

The previous chapters examined the production of intonational tunes in GuanM, first in single phrases in isolation (Chapter 3) and then in longer utterances in interaction with focus (Chapter 4). While these production studies revealed how question intonation is acoustically realised and how it interacts with lexical tone, they do not address whether listeners can reliably perceive and discriminate these intonational differences, nor whether lexical tones affect perception. Therefore, a perception study is necessary to establish whether the intonational patterns identified in production are communicatively functional. The present perception study focuses on single phrases in isolation, corresponding to the production data examined in Chapter 3. This chapter aims to determine how native listeners resolve the potential conflict between tone-dependent F0 movements and the question intonation. By employing a same-different (AX) paradigm, the study addresses two primary research questions:

- To what extent do listeners rely on the high boundary tone to identify question intonation when it is superimposed on different lexical tones?
- Does the phonetic interaction between tone and tune lead to a decrease in discrimination sensitivity *d*' in specific tonal contexts?

This chapter is organised into the following sections. The first section reviews the literature on the perception of intonation in Chinese languages. The second section presents the perception experiment, including the research questions, methods, and results (sensitivity scores, reaction times, and accuracy). The chapter concludes with a discussion of findings and conclusions.

5.1 Perception Studies in Chinese Languages

A central question in perception research focuses on identifying the main perceptual cues that enable listeners to identify intonation across different languages. These cues include pitch-related features such as pitch height and contour shape. Gussenhoven and Chen (2000) explored both universal and language-specific effects in the perception of question intonation by examining how listeners from three language backgrounds, Dutch, Hungarian, and Mandarin, perceived questions based on manipulated pitch patterns. They altered three key prosodic features, namely peak height, peak alignment, and end pitch, using trisyllabic stimuli from a pseudo-language to avoid familiarity. The results revealed that, across all the language groups, listeners tended to associate higher pitch peaks, later peaks, and higher end pitch with questions, suggesting that certain intonational cues, such as high pitch, are universally perceived as indicating questions. However, language-specific differences emerged; for instance, Hungarian listeners were more sensitive to peak height than were Mandarin listeners, indicating that the first language (L1) influences how intonational cues are processed.

Shang et al. (2024) examined how L1 Spanish speakers and Chinese learners of Spanish perceived yes-no questions and statements in Spanish. They found that both groups relied on final pitch movement to identify questions, indicating a universal prosodic feature in intonation perception. However, the Spanish speakers also relied on prenuclear peaks and stress patterns to differentiate between questions and statements, while Chinese speakers had more difficulty with intonational cues such as prenuclear peaks.

Perception studies of Chinese languages have paid particular attention to how lexical tone interacts with intonation in terms of pitch. This has led to ongoing debates regarding whether the primary perceptual cues are in the F0 height or the F0 contours of lexical tones, such as a higher F0 leading to the easier identification of questions in Chinese. Moreover, from another

perspective related to the temporal point F0 cues, researchers questioned whether the final prosodic word of an utterance, the last two syllables, or a more global tendency across the utterance, contributed to intonation perception.

Yuan (2011) examined how lexical tones in Mandarin affected L1 speakers' perceptions of intonation, particularly when differentiating between statements and questions. The results revealed that sentences ending with a falling tone (Tone 4) made question intonation easier to identify whereas those ending with a rising tone (Tone 2) made it more difficult, which indicated that the identification of tones played an active role in how pitch contours were mapped onto intonational categories, suggesting an interaction between tone and intonation at the phonological level. The study's findings also revealed asymmetry between the two types of intonation: Statement intonation was easier to identify and served as the default or unmarked form, while question intonation, being a marked type, was more frequently misidentified as a statement when the tonal or prosodic cues were unclear.

Similar findings were reported by B. R. Xu and Mok (2012), who observed that Mandarin speakers exhibited a perceptual asymmetry: Speakers could identify questions ending in Tone 4 more accurately than questions ending in Tone 2, both in complete sentences and those in which the last syllable had been omitted. Furthermore, the fact that question identification remained accurate even when the final syllable was missing suggested that Mandarin speakers relied on a global F0 height as a cue for questions. For Cantonese speakers, while the final syllable was a crucial cue, listeners still used pitch register rather than pitch contour to identify intonation when the last syllable was missing. This indicated that pitch height could sometimes take precedence over the contour shape. They further illustrated that both Mandarin and Cantonese listeners relied on universal pitch cues, such as higher pitch and rising contours, to

identify questions, even when listening to unfamiliar languages. J. K.-Y. Ma et al. (2006b) found similar results; in their study, perceptual accuracy between the full sentence condition and the final-syllable-only condition was similar, but was significantly lower when only the carrier sentence (without the final syllable) was presented, which suggested that the final rise in F0, which typically marked questions, was a crucial cue for intonation identification, whereas pitch changes earlier in the sentence (the carrier) were less effective. Furthermore, questions ending with high-rising tones were perceived more accurately than those ending with low-rising or other tones. By contrast, statements ending with high-rising tones were often misperceived as questions, demonstrating a tone and intonation interaction. In the follow-up study, J. K.-Y. Ma et al. (2011) discovered that questions and statements were distinguished by the F0 interval (the range of pitch between the onset and offset at the utterance-final syllable) and the average F0, both of which were significant predictors of question identification. By contrast, F0 slope (the rate of F0 change over time) was not found to be a significant predictor.

C. Zhang (2018) investigated the interaction between lexical tones and intonational tunes in Tianjin Mandarin using tune identification tasks and revealed that listeners rely on two primary cues when distinguishing intonational yes-no questions from statements: register information and a floating H% boundary tone. Crucially, lexical tones were found to compete with intonational cues during perception; high-rising tones facilitated question identification, whilst low-falling tones facilitated statement identification. When boundary information was removed from stimuli, identification accuracy dropped markedly, emphasising the importance of the boundary tone in question perception. Furthermore, reaction time data provided evidence for incremental processing, suggesting that listeners begin formulating tune judgements upon encountering early register cues rather than awaiting complete utterances.

Liang and van Heuven (2007) explored how L1 Mandarin speakers and learners of Mandarin as a second language (L2), whose L1s were either tonal (Nantong and Changsha dialects) or nontonal (Uygur), perceived both lexical tones and sentence intonation using global rising or final-rising cues. By manipulating the final rise in pitch, the authors found that sensitivity to the final pitch movement was a crucial determinant in listeners' judgements when compared to the overall pitch level. Of note, nontonal-language listeners identified sentence types more quickly and were more responsive to sentence-level pitch cues. However, L1 Mandarin speakers were slower, which was probably due to the cognitive demands of processing both lexical tones and intonation simultaneously. Such results indicated a trade-off between tone and intonation processing whereby speakers of tonal languages tended to prioritise lexical tone, thus decreasing their sensitivity to sentence-level intonation cues.

Ni and Kawai (2004) investigated how changes in pitch targets (F0 peaks and valleys) influenced the perception of tone and intonation in Mandarin. They modified the pitch targets of the final tones in two carrier sentences, which were originally statements, to observe how these changes affected listeners' perceptions of the utterances as statements or as questions, as well as the perceived focus and tone of the final syllable, by either keeping the F0 peak or valley unchanged and altering the other, or by changing both peak and valley. The results revealed that the listeners consistently perceived utterances with higher F0 peaks as being questions. Specifically, for sentences ending in Tone 2, higher F0 peaks made the sentence more likely to be interpreted as a question. However, for sentences ending in Tone 4, the listeners only perceived the utterance as being a question when the entire tone was positioned in a higher pitch register. If the F0 valley remained low, the utterance was always judged to be a statement.

Liu et al.'s (2022) study examined how semantic context influenced the processing of tone and intonation in Mandarin, particularly when the pitch contours of the lexical tone and the sentence intonation were either aligned or not aligned, as in the case of Tone 2 and Tone 4. The authors compared how Mandarin speakers identified tone and intonation in both semantically neutral and semantically constraining contexts. Here, only the results of the semantically neutral contexts from their paper are presented; the carrier sentence used was "She just said X", with the final syllable X varying in tone (T2 or T4) and produced with either statement or question intonation. In this condition, the participants remained highly accurate in terms of tone identification even in the absence of semantic cues, which suggested that the listeners relied heavily on acoustic features to differentiate between tones. However, identifying intonation was much more challenging in the neutral context, particularly for questions. The participants found it equally difficult to differentiate between question and statement intonation when the final tone was either T2 or T4, as the lack of semantic information (top-down information) made it more challenging to separate the inherent pitch contour of the tone and the intonational pattern.

Yang et al. (2024) explored another aspect of tone and intonation by investigating whether the perception of intonation was gradient or discrete in two varieties of Chinese: Cantonese, which has lexical tones that differ in both pitch register and pitch shape, and Zhumadian Mandarin (ZM), a dialect with four lexical tones, namely an early fall, a late fall, an early rise, and a late rise. In Cantonese, questions are marked by a final pitch rise, while questions are expressed through the raising and expansion of pitch range in ZM. The study's findings revealed that lexical tone contrasts were perceived categorically in both Cantonese and ZM, in line with the established view that tonal contrasts are phonologically encoded. However, with regard to the perception of intonation, in Cantonese, in which intonation is marked by pitch shape

differences, intonation contrasts were also perceived categorically. By contrast, ZM speakers, whose intonation contrasts rely on the pitch height factor, perceived these contrasts gradiently. These results indicated that the phonological encoding of pitch shape was essential for categorical perception, while pitch heights linked to intonational significance were processed in a more continuous manner.

In summary, these studies of Chinese languages have validated the complex interaction between lexical tone and intonation in perception by showing that both F0 height and contour influence how listeners perceive intonation. While final syllable cues, such as pitch rise or fall, are important for differentiating between statements and questions, global pitch patterns also play a crucial role; which factor serves as the primary or secondary cue in perception depends on the language. In addition, the cognitive load involved in processing both tone and intonation can vary between speakers of tonal and nontonal languages, with speakers of tonal languages often prioritising lexical tone, which may reduce their sensitivity to sentence-level intonational cues.

5.2 Methods

5.2.1 Participants

Forty native speakers of GuanM, including 20 females from Xi'an, China, participated the experiment (mean age: 36.95 years, age range 20 to 50 years). None of them reported having hearing or speaking disorders. Before taking part in the study, all the participants gave written informed consent and were compensated for their time.

5.2.2 Stimuli

The stimuli comprised recordings obtained in Production Study 1 and consisted of 36 disyllabic words: 12 each for the T1T1, T2T2, and T4T4 (see Table 5.1 for the full list of words). The items had been produced by a single 24-year-old female speaker of GuanM, who recorded each word in both question and statement intonation. The mean duration of the recorded stimuli was 757.37 ms for questions and 741.66 ms for statements.

Table 5.1 List of stimuli used in the Perception Study

T1T1	**T2T2**	**T4T4**
乌鸦	麻油	浪漫
u31ia31	ma24iou24	laŋ55man55
'crow'	'sesame oil'	'romance'
鸳鸯	绵羊	面料
yan31iaŋ31	mian24iaŋ24	mian55liau55
'mandarin duck'	'sheep'	'fabric'
落叶	棉麻	面貌
luo31iɛ31	mian24ma24	mian55mau55
'fallen leaves'	'linen fabric'	'face'
音乐	蓝莓	样貌
ien31io31	lan24mei24	iaŋ55mau55
'music'	'blueberry'	'appearance'
录音	羊毛	外卖
lu31ien31	iaŋ24mau24	uæ55mæ55
'recording'	'wool'	'takeout'
弯月	农民	命运
uan31yɛ31	nuŋ24mien24	miŋ55yen55
'crescent moon'	'farmer'	'destiny'
阅历	南门	外贸
yɛ31li31	nan24men24	uæ55mau55
'experience'	'south gate'	'foreign trade'
月末	杨梅	命令
yɛ31mo31	iaŋ24mei24	miŋ55liŋ55
'end of month'	'bayberry'	'order'
目录	毛驴	另类
mu31lu31	mau24ly24	liŋ55luei55
'catalogue'	'donkey'	'alternative'

优越 iou31yɛ31 ‘superior’	煤油 mei24iou24 ‘kerosene’	暧昧 ŋæ55 mei55 ‘ambiguous’
沐浴 mu31y31 ‘bathing’	邮轮 iou24lyen24 ‘cruise’	命案 miŋ55ŋan55 ‘homicide’
烟叶 ian31iɛ31 ‘tobacco’	牦牛 mau24ŋiou24 ‘yak’	议论 i55lyen55 ‘discussion’
邀约 iau31ŋio31 ‘invite’	鱿鱼 iou24y24 ‘squid’	义卖 i55mæ55 ‘charity sale’

5.2.3 Procedure

Research on speech perception typically uses two main categories of tasks: identification tasks and discrimination tasks (H.-S. Cheng et al., 2021). For this experiment, the AX discrimination task was selected over both identification tasks and other discrimination paradigms such as ABX or 4I2AFC (four-interval, two-alternative forced-choice). In the AX task, participants are presented with two successive sounds (A and X) and judge whether the sounds are the same or different. By contrast, the ABX task requires participants to hear three sounds (A, B, and X) and to decide whether the third sound (X) matches the first (A) or the second (B), which increases the cognitive load and encourages labelling strategies that are based on phoneme categories rather than on purely auditory discrimination. The 4I2AFC task presents participants with a sequence of sounds in either the AABA or the ABAA order, with stimulus A at the beginning and at the end functioning as a reference. Participants must determine whether the odd sound occurs in the second or third interval. This task involves both auditory processing and phoneme labelling, but requires more cognitive effort due to the need to track multiple intervals and sounds (Gerrits & Schouten, 2004). Since the focus of the task was on the perception of intonation in tones, in which pitch serves as the primary cue and there is already complicated interaction between intonation and tone, the AX task was chosen to decrease the

cognitive load via the experimental design itself. This allowed the participants to focus more on auditory perception without the interference of additional cognitive demands.

The decision to omit a formal identification task was strategically made to prioritise the measurement of raw auditory sensitivity over linguistic labelling. In identification tasks, listeners explicitly label a sound as belonging to one category or another, which taps into a categorical mode of perception (H.-S. Cheng et al., 2021). This forced categorisation may mask the gradient nature of intonation contrasts. This distinction is critical, as Gussenhoven and Van de Ven (2020) demonstrated in their study of Zhumadian Mandarin: while lexical tone contrasts (early vs. late alignment of falls and rises) were perceived categorically, the statement-question intonation contrast was perceived gradiently. The AX discrimination task allows for the detection of sensitivity to acoustic variations within the same category, effectively focusing on auditory pitch cues, such as pitch range and alignment, without requiring categorical labelling. Therefore, the AX discrimination task was used in this experiment to assess the participants' ability to differentiate between different intonations of the same word. Each trial had two sounds: Sound A (the first sound) and Sound X (the second sound), both of which were the same word produced with either a statement or a question intonation with possible combinations based on the order of the sounds:

A = Statement, X = Statement (SS)

A = Statement, X = Question (SQ)

A = Question, X = Question (QQ)

A = Question, X = Statement (QS)

Each participant was presented with these four combinations twice, resulting in a total of 288 trials per participant (3 tones x 12 disyllabic words x 4 combinations x 2 repetitions). The order

of the presentation was randomised to minimise order effects and to enhance the reliability of the responses.

The experiment was conducted using PsychoPy (Peirce et al., 2019). The participants were seated in a quiet room in a local community centre and wore headphones. They were instructed to listen carefully to each pair of sounds and to decide whether the intonation of the second sound (X) matched that of the first sound (A) by pressing "1" for "same" or "0" for "different". They were given 2000 ms to respond, with a 200 ms interval between the two sounds. The entire experiment took approximately 15 minutes to complete.

5.2.4 Statistical Analyses

To assess the participants' ability to differentiate between different pairs of sounds, the *d*-prime (*d'*) score was calculated for each participant based on each tone. These scores were divided into two pairs based on the intonation of the first sound (A): SS_SQ (when A = Statement) and QQ_QS (when A = Question), which was used to explore whether the sequence of intonations made a difference in the AX task. The following formula was used to compute *d'*:

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

The hit rate is the percentage of the times that the participants correctly identified the second sound (X) as being different from the first sound (A) when there was an actual difference. For example, in the SQ and QS conditions, where A = Statement and X = Question (or vice versa), the participants should correctly perceive the difference in intonation. The hit rates for SQ and QS were calculated by dividing the correct responses in each condition by the total number of trials for SQ and QS, respectively. The false alarm rate refers to the percentage of times that

the participants incorrectly identified the second sound as being different when they were actually the same, such as in SS and QQ conditions, in which both sounds (A and X) had the same intonation. The false alarm rates for SS and QQ were calculated by dividing the number of incorrect responses for each condition by the total number of trials for SS and QQ, respectively. The *d'* score is the difference between the Z-scores for the hit rate and the false alarm rate. A higher *d'* value indicates greater sensitivity or discrimination ability (Macmillan & Creelman, 2004), meaning that the participants were more accurate in detecting when the two sounds differed without having many false alarms.

To analyse the *d'* scores, a linear mixed-effects model was fitted using the lmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2024), where the *d'* score was the dependent variable, with Pair (SS_SQ and QQ_QS) and Tone (T1T1, T2T2, and T4T4) and their interaction as fixed effects, and Participant as a random effect.

Reaction times (RTs) for correct responses were also analysed. Outliers were removed prior to analysis: RTs more than three standard deviations above or below each participant's mean were excluded, resulting in a data loss of 1.48% (170 trials). The RTs were log-transformed to reduce the positive skew typical of RT distributions, following common practice. A linear mixed-effects model was fitted with RT as the dependent variable. The fixed effects were Tone (T1T1, T2T2, and T4T4), AX Sequence (SS, SQ, QQ, and QS) and their interaction. Participant, Item and Repetition were included as random effects.

The accuracy data were analysed using a generalised linear mixed-effects model. Both the fixed and the random effects in the accuracy model were the same as those in the RT model. Post-hoc pairwise comparisons for significant effects were conducted using Tukey's HSD tests in

the emmeans package (Lenth, 2024) in R (R Core Team, 2024). The model was specified in R as follows:

model_accuracy <- glmer(Accuracy ~ Tone* AX_Sequence + (1 | Participant) + (1 | Item) + (1 | Repetition), data = df, family = binomial)

5.3 Results

5.3.1 *d'* Scores

Figure 5.1 presents the *d'* scores for T1T1, T2T2, and T4T4 across two Pair pairs: SS_SQ and QQ_QS. As shown, the SS_QQ pair had an overall higher *d'* score compared to the QQ_QS pair, suggesting that the participants were better at distinguishing the difference when the first tone was statement intonation. T1T1 consistently showed the highest sensitivity across both pairs. When the first sound was a statement, the *d'* score was 2.51, compared to 2.34 when the first sound was a question. By contrast, T2T2 demonstrated the lowest sensitivity in the QQ_QS pair as the mean *d'* score was only 1.00, suggesting that the participants had more difficulty detecting the difference in T2T2 when the first intonation was question intonation. However, T2T2 showed a significant improvement in the SS_SQ pair with 2.15 *d'* scores, indicating that the participants found it easier to perceive the differences in two intonations when the first sound had statement intonation. T4T4's performance was intermediate between T1T1 and T2T2, with a mean *d'* score of 1.64 in the QQ_QS pair and 1.92 in the SS_SQ pair.

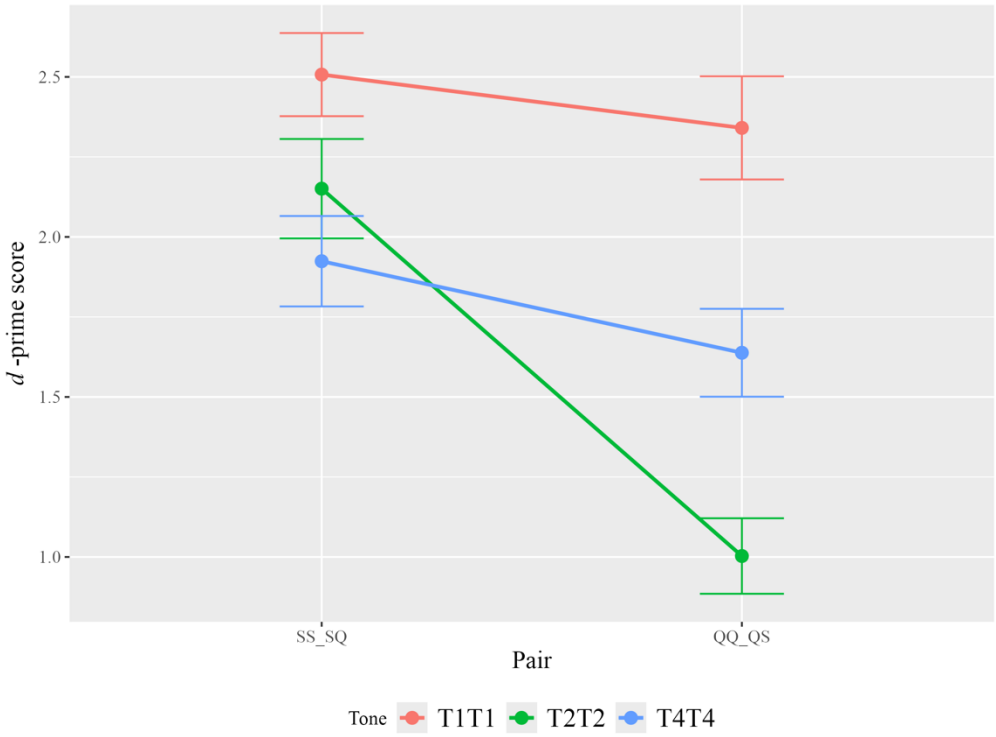


Figure 5.1 *d'* scores for T1T1, T2T2 and T4T4 by Pair: SS_SQ and QQ_QS

The results of the statistical analysis revealed that the main effects of Tone, Pair, and their interaction were all significant in terms of *d'* scores (Tone: $\chi^2(4) = 150.53$, $p < .001$; Pair: $\chi^2(3) = 111.89$, $p < .001$; Tone x Pair: $\chi^2(2) = 52.65$, $p < .001$). The results of the post-hoc pairwise comparisons showed that, for the QQ_QS pair, when the first sound had a question intonation, the *d'* score for T1T1 was significantly higher than it was for T2T2 ($\beta = 1.34$, $t(195) = 13.54$, $p < .0001$) and higher than it was for T4T4 ($\beta = 0.70$, $t(195) = 7.11$, $p < .0001$). The *d'* score for T2T2 was significantly lower than it was for T4T4 ($\beta = 0.64$, $t(195) = -6.43$, $p < .0001$). The differences between the tones were less distinct for the SS_SQ pair when the first sound had a statement intonation. T1T1's *d'* score was still significantly higher than were the scores for T2T2 ($\beta = 0.36$, $t(195) = 3.61$, $p = .005$) and T4T4 ($\beta = 0.58$, $t(195) = 5.90$, $p < .0001$), although no significant difference was found between T2T2 and T4T4 ($p = .201$). Furthermore, both T2T2 and T4T4 had significantly increased *d'* scores from the QQ_QS to the SS_SQ pair, with

T2T2 showing a larger increase (T2T2: β = -1.15, t(195) = -11.62, p < .0001; T4T4: β = -0.29, t(195) = -2.90, p = .048). This suggested that the ability to discriminate between T2T2, a rising tone, and T4T4, a high tone, was enhanced when the first tone was a statement, which was not the case for T1T1 in the QQ_QS and SS_SQ pairs (β = -0.17, t(195) = -1.69, p = .543). See Table 5.2 for the output of the model for the *d'* scores.

Table 5.2 Model's output for the *d'* scores. Square brackets indicate the factor level comparisons

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.34	2.06 – 2.62	**<.001**
Tone [T2T2]	-1.34	-1.53 – -1.14	**<.001**
Tone [T4T4]	-0.70	-0.90 – -0.51	**<.001**
Pair [SS_SQ]	0.17	-0.03 – 0.36	.093
Tone [T2T2] × Pair [SS_SQ]	0.98	0.71 – 1.26	**<.001**
Tone [T4T4] × Pair [SS_SQ]	0.12	-0.16 – 0.39	.393
Random Effects			
σ^2	0.20		
$\tau_{00\ Participant}$	0.60		
ICC	0.76		
$N_{\ Participant}$	40		
Observations	240		
Marginal R^2 / Conditional R^2	0.239 /0.814		

5.3.2 RTs

Significant main effects were observed for both AX Sequence (χ^2(9) = 29.74, p < .001) and Tone (χ^2(8) = 18.93, p = .015), showing that both AX Sequence and Tone had a significant impact on the RTs, while the interaction effect between the AX Sequence and Tone was not significant (χ^2(6) = 12.42, p = .053). Post-hoc comparisons using Tukey's test further revealed

that the significant differences across the sequences only occurred for T1T1. For T1T1, the RTs for the SS sequence were significantly longer than were those for the SQ sequence (β = 0.16, t(7956) = 3.10, p = .0106); the RTs for the SQ sequence were also significantly shorter than were those for the QQ (β = -0.14, t(7965) = -2.71, p = .0341) and QS sequences (β = -0.22, t(7953) = -4.03, p = .0003). No other significant effects were found. No significant differences were observed between sequences for T2T2, although the trend suggested faster responses when the first tone was a statement. Similarly, T4T4 showed no significant differences across sequences, although a similar trend of faster responses for the sequence in which the first sound had statement intonation was observed. See Table 5.3 for the detailed mean RTs and Table 5.4 for the summary of the statistical result.

Table 5.3 RTs for correct responses across different AX Sequences and Tones

AX Sequence	Tone	N	Mean (SD)
A = Statement, X = Statement	T1T1	759	369.55 (293.19)
	T2T2	737	364.31 (307.87)
	T4T4	705	375.06 (316.73)
A = Statement, X = Question	T1T1	701	346.93 (314.61)
	T2T2	616	368.11 (314.29)
	T4T4	615	376.10 (301.11)
A = Question, X = Question	T1T1	758	364.20 (288.41)
	T2T2	663	395.00 (329.39)
	T4T4	747	375.59 (302.83)
A = Question, X = Statement	T1T1	701	389.09 (324.12)
	T2T2	478	375.90 (299.69)
	T4T4	543	389.00 (311.12)

Table 5.4 Model's output for log-transformed RTs. Square brackets indicate the factor level comparisons

Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.49	5.33 – 5.64	**<.001**

AX Sequence [A=S, X=Q]	-0.16	-0.27 – -0.06	**.002**
AX Sequence [A=Q, X=Q]	-0.02	-0.12 – 0.08	.693
AX Sequence [A=Q, X=S]	0.05	-0.05 – 0.16	.314
Tone [T2T2]	-0.08	-0.18 – 0.03	.148
Tone [T4T4]	-0.01	-0.12 – 0.10	.874
AX Sequence [A=S, X=Q] × Tone [T2T2]	0.19	0.04 – 0.33	**.014**
AX Sequence [A=Q, X=Q] × Tone [T2T2]	0.13	-0.02 – 0.27	.082
AX Sequence [A=Q, X=S] × Tone [T2T2]	0.09	-0.06 – 0.25	.241
AX Sequence [A=S, X=Q] × Tone [T4T4]	0.23	0.08 – 0.38	**.002**
AX Sequence [A=Q, X=Q] × Tone [T4T4]	0.08	-0.07 – 0.22	.290
AX Sequence [A=Q, X=S] × Tone [T4T4]	0.05	-0.10 – 0.20	.524
Random Effects			
σ^2	1.00		
$\tau_{00\ Participant}$	0.17		
$\tau_{00\ Item}$	0.00		
$\tau_{00\ Repetition}$	0.00		
ICC	0.15		
$N_{\ Participant}$	40		
$N_{\ Item}$	36		
$N_{\ Repetition}$	2		
Observations	8023		
Marginal R^2 / Conditional R^2	0.004 / 0.151		

5.3.3 Accuracy

As shown in Figure 5.2, T1T1 consistently showed the highest accuracy across all the sequences, particularly in the SS (80.06%) and QQ (80.13%) sequences. By contrast, T2T2 consistently showed lower accuracy compared to both T1T1 and T4T4. For example, in the QS sequence, T2T2 had the lowest accuracy at 50.80% which was at chance level (50%), suggesting participants performed no better than guessing. T4T4 had variable results: higher

than T2T2 in SS and QQ sequences, but showing similar accuracy to T2T2 in the QS sequence. When examining the effects of the AX Sequence, the SS and QQ sequences exhibited the highest accuracy across all tones, with T1T1 being the highest. However, the SQ sequence showed a decline in accuracy compared to the SS, particularly for T1T1 (64.84%) and T2T2 (65.36%). The QS sequence had the lowest overall accuracy, particularly for T2T2 (50.80%) and T4T4 (57.70%).

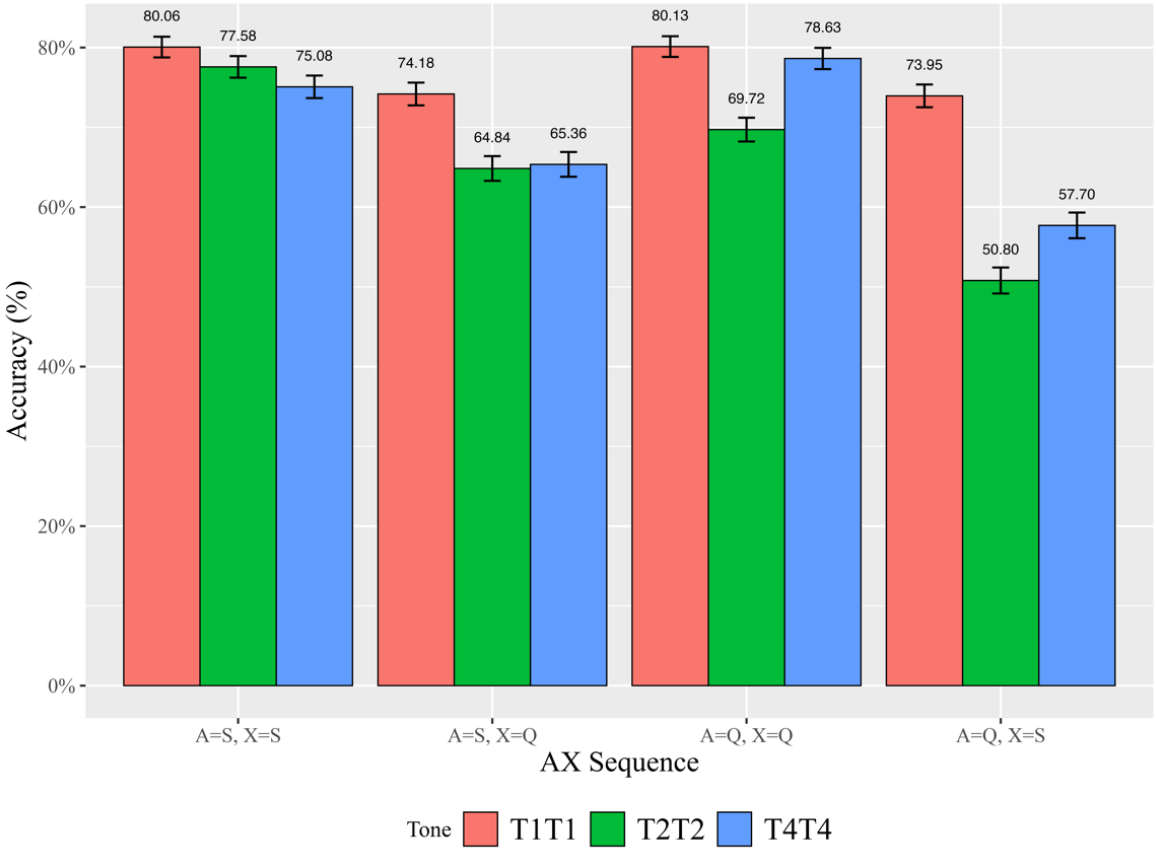


Figure 5.2 Accuracy across different AX Sequences and Tones

The statistical results showed that, for accuracy, the main effects of the AX Sequence ($\chi^2(9) = 357.73$, $p < .001$) and Tone ($\chi^2(8) = 87.51$, $p < .001$) and the interaction between the AX Sequence and Tone were all significant ($\chi^2(6) = 61.85$, $p < .001$), indicating that the effect of

the AX Sequence on accuracy depended on the level of Tone. The observations based on Figure 5.2 were further confirmed by the Tukey's HSD pairwise comparisons. In the SS sequence, T1T1 had higher accuracy than both T2T2 (β = 0.17, z = 1.08, p = .529) and T4T4 (β = 0.32, z = 2.07, p = .096), although these differences were not statistically significant. However, in the SQ sequence, T1T1's accuracy was significantly higher than both T2T2 (β = 0.54, z = 3.59, p = .001) and T4T4 (β = 0.50, z = 3.34, p = .003). In the QQ sequence, the participants continued to respond significantly better in T1T1 than in T2T2 (β = 0.66, z = 4.30, p = .0001), but no significant difference was found between T1T1 and T4T4. In addition, T4T4 showed significantly higher accuracy than did T2T2 (β = -0.56, z = -3.63, p = .0008). In the QS sequence, T1T1 maintained higher accuracy than both T2T2 (β = 1.23, z = 8.32, p < .0001) and T4T4 (β = 0.88, z = 5.95, p < .0001), with T4T4 also showing significantly higher accuracy than T2T2 (β = -0.35, z = -2.43, p = .0403). In sum, T1T1 consistently achieved the highest accuracy across sequences, whereas accuracy varied markedly across AX Sequences: it remained high in the SS and QQ sequences but dropped substantially in the SQ and, especially, QS sequences. T2T2 was the most difficult for participants to distinguish, particularly when the sequence involved a shift from question to statement intonation (QS). See Table 5.5 for the summary of the statistical model's output for accuracy.

Table 5.5 Model's output for Accuracy. Square brackets indicate the factor level comparisons

Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	5.36	3.65 – 7.87	**<.001**
AX Sequence [A=S, X=Q]	0.67	0.53 – 0.85	**.001**
AX Sequence [A=Q, X=Q]	1.01	0.79 – 1.29	.938
AX Sequence [A=Q, X=S]	0.67	0.53 – 0.84	**.001**
Tone [T2T2]	0.84	0.62 – 1.15	.282
Tone [T4T4]	0.72	0.53 – 0.98	**.038**

AX Sequence [A=S, X=Q] × Tone [T2T2]	0.69	0.50 – 0.96	**.025**
AX Sequence [A=Q, X=Q] × Tone [T2T2]	0.61	0.44 – 0.85	**.003**
AX Sequence [A=Q, X=S] × Tone [T2T2]	0.34	0.25 – 0.47	**<.001**
AX Sequence [A=S, X=Q] × Tone [T4T4]	0.84	0.61 – 1.16	.284
AX Sequence [A=Q, X=Q] × Tone [T4T4]	1.24	0.89 – 1.74	.203
AX Sequence [A=Q, X=S] × Tone [T4T4]	0.57	0.42 – 0.79	**.001**
Random Effects			
σ^2	3.29		
$\tau_{00\ Participant}$	1.00		
$\tau_{00\ Item}$	0.06		
$\tau_{00\ Repetition}$	0.00		
ICC	0.24		
$N_{Participant}$	40		
N_{Item}	36		
$N_{Repetition}$	2		
Observations	11350		
Marginal R^2 / Conditional R^2	0.055 / 0.286		

5.4 Discussion

Tune perception relies on tone-dependent tonal interactions. In the case of T1T1 (low falling tone), the first syllable undergoes lexical tone dissimilation, changing from a falling tone to a rising tone, while the second syllable remains falling. However, when a question tune is present, the second syllable's falling tendency is reduced; this means that it does not fall as sharply as it would with a statement tune due to the high boundary tone of the question. The interaction between the higher register of the question tune and the boundary tone causes the second syllable not to fall as much, thus lifting the entire tone's disyllabic contour upwards.

In question tunes, both T1T1 and T2T2 have faster-rising first syllables T4T4 show a slight rise due to the influence of the high boundary tone and the overall higher register of the question tune. As for the second syllable, T1T1's second syllable remained a slightly falling tone, which was incongruent with the rising contour of the question tune. By contrast, T2T2's second syllable, being a fast-rising tone, aligned well with the rising contour of the question, while T4T4's slight rise in the second syllable made it almost congruent with the question tune contour. The *d'* score results suggest that the final syllable in GuanM is a crucial cue for perception. T1T1's incongruent final syllable produced the highest *d'* scores across both pairs, regardless of whether the first sound was a statement or a question, indicating a stronger ability to distinguish between question and statement tunes compared to T4T4 and T2T2, which had the lowest *d'* scores when the first sound was a question. These findings are consistent with previous research on Mandarin, in which T2 (a rising tone) is known to be more challenging to identify, while T4 (a falling tone) is generally easier to identify. A further analysis of the *d'* score results revealed that the scores were significantly lower for the QQ_QS pair than for the SS_SQ pair for both T2T2 and T4T4, with the difference for T2T2 being greater; this suggests that the participants had difficulty differentiating between question and statement tunes when the first sound that they heard was a rising tone or a high-level tone in question intonation, with the rising tone causing more perceptual difficulty.

Regarding accuracy and RTs, T1T1 consistently exhibited strong performance. The participants had higher accuracy, even in slower response conditions such as the QQ sequence, suggesting that they took more time to ensure precision. Nonetheless, the accuracy dropped in quicker response situations such as the SQ sequence, indicating a clear trade-off between RTs and accuracy. On the contrary, T2T2 showed slower RTs and less accuracy, particularly in the QS sequence for which accuracy fell to 50.8%, reflecting a struggle to process this tone when

the first sound was a question. The results were mixed for T4T4. The participants responded well when the tune matched, as in the SS and QQ sequences, but had difficulties in the QS sequence, for which the accuracy decreased significantly. The longer RTs that were observed for T4T4 imply that the participants needed more time to process this tone, although it was not as challenging as T2T2.

When examining the AX Sequence, the results show that the participants performed better when both sounds had matching intonations. The SS sequence was marked by high accuracy and relatively fast responses, meaning it is less difficult to process. By contrast, the QQ sequence had slower RTs but higher accuracy, suggesting that the participants took extra time to ensure accuracy, particularly with T1T1 and T4T4. However, performances generally declined when the intonations of the two sounds differed. There was a clear speed-accuracy trade-off in the SQ sequence, particularly for T1T1 and T2T2, as faster responses were associated with less accuracy. The QS sequence proved to be the most difficult, as it was characterised by both slower RTs and less accuracy. This was particularly true for T2T2, as the accuracy dropped to 50.8%, and for T4T4, which had an accuracy of 57.7%, reflecting the difficulty of processing a question followed by a statement.

These findings further confirm that tune information interacts with tone information during perception. When the two sounds had different tunes, the participants found it difficult to separate the tune and the tone. Perception became more difficult when the tone contour aligned with the tune contour, as in T2T2, but the task was easier when the contours did not align, as in T1T1. The relatively lower accuracy for T4T4 in the QS sequence could be attributed to the higher register of the question tune, which may have made it more difficult for the participants to identify the intonation. This potentially suggests that the higher register could also serve as

a cue for the perception of a question tune in GuanM. As Gussenhoven and Chen (2000) argued, listeners tend to associate high pitch and rising contours with questions, which might complicate tone perception in certain contexts.

In addition, the increased RTs noted in the QQ and QS sequences suggest that the task of storing, retrieving, and comparing a question tune with a following tune demands considerable cognitive effort and memory processing, resulting in delayed responses. In comparison, the RTs were quicker in the SS and SQ sequences, where the initial sound was a declaration. This could be attributed to the statement tune functioning as a default, thus requiring less cognitive effort to remember, access, and contrast with the subsequent tune.

In summary, when a question tune is first triggered and held in short-term memory, it is very challenging to retrieve and align it with the subsequent statement or question. This difficulty arises from the need to retrieve additional information about the question tune, including its contour and register. The relationship between lexical tone and tune shape is crucial in this process. When the tone contour does not match the tune contour, as in T1T1, perception remains very straightforward. However, when the shapes match, as seen in the case of T2T2, the perceptual and cognitive demands increase, which leads to longer RTs and lower accuracy.

5.5 Conclusion

In conclusion, this pilot study explored the perception of yes-no question tunes in GuanM by focusing on the interaction between tune and tone using the AX paradigm. The findings revealed two key points: First, the final syllable plays a critical role in intonational perception. Second, the study demonstrated the complex interaction between tone and tune during perception, particularly regarding contour shape, thus showing that tone-dependent factors

significantly influence how question and statement tunes are processed. This is where the distinction of perception between question and statement tunes becomes most obvious.

The results revealed that the participants were better able to distinguish between tunes when the tone contour of the final syllable did not align with the tune contour, as seen in T1T1, leading to higher accuracy and faster RTs. Conversely, perception became more challenging when the tone contour aligned with the tune contour, as in T2T2, resulting in slower RTs and reduced accuracy. In addition, the findings suggest that the higher register of a question tune adds another layer of difficulty, particularly for tones such as T4T4, which are less congruent with rising contours. The observed patterns of longer RTs in question sequences and speed-accuracy trade-offs emphasise the cognitive effort that is required to process mismatched tunes, particularly when a question tune is presented first. The findings also suggest that the storage and retrieval of question tunes require greater working memory load and cognitive processing due to the higher register and high boundary tone associated with question intonation, whereas the statement tune acts as a default.

Despite the initial intention of using the AX task to reduce cognitive load, its limitations should be acknowledged. The task may introduce bias, as the participants may only select “different” when they are highly confident about their judgements, potentially skewing the results because responses may be heavily influenced by a subjective phoneme-based criterion, leading participants to rely on a decision that falls near one extreme of the “same” versus “different” scale, thus affecting the overall interpretation of the data (Gerrits & Schouten, 2004).

Chapter 6 Planning of Tunes in Guanzhong Mandarin

The previous chapters examined both the production of intonational tunes and the perception of tunes in GuanM. The findings confirmed that during perception, tone and tune interact, with the nature of the interaction depending on whether the contour shapes of tone and tune are congruent or incongruent. This chapter describes an investigation of the planning of tunes using the picture-word interference (PWI) paradigm from a psycholinguistic perspective. This pilot study aims to explore how tunes, particularly the question tune, are planned during production in GuanM while audio distractors, either semantic or phonological, are introduced at three different stimulus onset asynchronies (SOAs). The chapter opens with a literature review of spoken language planning and intonation planning. Following this, the experimental method using the PWI paradigm is discussed in detail, including participant criteria, experimental procedure, and methods used for the statistical analysis. Results are then presented, focusing on tune effects and tone interaction during the planning stage. It concludes with an interpretation of the findings, which suggest that tune planning occurs at a very early stage in production, as well as a discussion of the study's limitations.

6.1.1 Word-level Planning

This section primarily addresses processing below the sentence level. Since the interaction of tone and intonation is highly complex and lexical tones carry both phonological and semantic-lexical information, it is essential to first examine how lexical tone is represented and processed during spoken word planning before investigating how tunes are planned at the sentence level.

First, tone is actively planned during speech production. Early speech-error studies in Mandarin and Thai reported tonal errors that closely mirror segmental errors in both form and distribution,

including anticipation, perseveration and exchange effects (Wan & Jaeger, 1998). More recently, large-scale corpus analyses of Cantonese conversational speech have shown that tonal errors are not only frequent but also predominantly contextual, which is more like consonant and vowel errors than like stress errors in English, indicating that tone is selected during phonological encoding rather than being assigned automatically at a later stage (Alderete et al., 2019).

However, there has been ongoing debate over whether tone is encoded early or late in the speech-planning process. Early-encoding accounts propose that tone is stored as part of a word's phonological representation and selected alongside segments. Wan (2007) analysed thousands of Mandarin speech errors and found that tone errors occur just as freely as segmental errors and are strongly influenced by neighbouring syllables, suggesting that tones are lexically specified and organised within the phonological system. In related work, Wan and Jaeger (1998) demonstrated that tone can be separated from segmental content in speech errors, further supporting its status as an independent representational unit.

In contrast, late-encoding accounts argue that tone functions more like metrical stress, being mapped onto prosodic frames rather than actively selected during lexical encoding. J.-Y. Chen (1999), for example, claimed that genuine tone errors are rare and that many apparent tonal slips are better explained as cases of lexical substitution. However, this position has been challenged by later work using larger and more naturalistic datasets, such as Alderete et al.(2019) who showed that tone errors are neither marginal nor accidental, but instead reflect systematic planning processes.

Experimental findings provide further explanations about how tone is planned. O'Seaghdha et al. (2010) found that Mandarin speakers benefit from advance knowledge of the first syllable, but not from knowledge of individual segments, suggesting that the syllable serves as a primary planning unit. Crucially, because tone is obligatorily associated with the syllable in Mandarin, this implies that tonal planning is tightly integrated with syllabic planning. Similarly, Wong and Chen (2008) used picture-word interference experiments in Cantonese and found facilitation effects when targets and distractors share both syllable and tone, demonstrating that tone plays a direct role in phonological encoding.

Computational modelling additionally supports the view of tones directly related to phonological encoding. Extensions of the WEAVER++ model to Mandarin suggest that, although languages differ in their basic phonological units, the core mechanisms of speech production remain broadly consistent. In Mandarin, atonal syllables are retrieved first and then associated with tonal frames, allowing tone to be integrated into phonological planning rather than relegated to phonetic implementation (Roelofs, 2015). While tone interacts with syllables and segments differently from how stress functions in English, the evidence strongly suggests that tone is a central component of phonological encoding in tonal languages.

6.1.2 Sentence Planning

Spoken language production requires the coordination of multiple cognitive stages, including conceptual activation, lexical retrieval, and phonological encoding. Central to this question are the discussions about serial and interactive models of lexical access, that is, whether semantic processing must be fully completed before phonological encoding begins, or whether these processes overlap and interact to some degree. The serial models propose a strictly sequential, two-stage process. The traditional explanation for the serial model was suggested by Levelt

(1992). In this two-phase model, speakers initially obtained a word's lemma that included its semantic and syntactic characteristics, and then the lexeme, which held the phonological representation. These phases were regarded as modular and lacking interaction. Supporting this view, Schriefers et al. (1990) found non-overlapping time windows between semantic and phonological effects using the PWI paradigm: semantic interference occurred when distractors were presented early (SOA = -150 ms), whereas phonologically related distractors speeded naming at later SOAs (0 ms or 150 ms).

Challenging the strict modularity assumed by serial models, several studies have provided evidence for cascade models that allow for continuous information flow between processing levels. For example, Humphreys et al. (1988) initiated a cascade model that had a fluid flow of information between levels of representation by showing that visual and semantic similarity influenced naming latencies in an overlapping manner in picture identification tasks. Damian and Martin (1999) found that semantic interference was reduced or eliminated when distractor words were also phonologically related to the target. Such results contradicted the additive predictions of serial models but supported the idea of bidirectional activation between semantic and phonological levels. Similarly, Morsella and Miozzo (2002) demonstrated that multiple lexical candidates could become phonologically activated. This result is incompatible with serial models, which predict that only selected words should influence phonological encoding. The phonological information from unselected words could influence naming by using a picture-picture interference paradigm, in which naming was faster when target pictures (e.g., "bed") were paired with phonologically related distractors (e.g., "bell"), than compared to unrelated distractors.

Schnur et al. (2006) advocated for non-modular models of speech production and extended the discussion of phonological planning by testing whether phonological encoding was confined to a single word at a time (a hypothesis known as radical incrementality). Their experiments showed that participants started speech faster when a phonologically related distractor matched a verb later in the sentence, even if that verb occurred as the second or third phonological word, suggesting phonological planning extended across multiple words and was not strictly limited by phrase boundaries. Wheeldon and Lahiri (2002) examined the minimal unit of phonological encoding in Dutch by comparing production latencies for compounds (e.g., "ooglid" meaning "eyelid"), adjective-noun phrases (e.g., "oud lid" meaning "old member"), and morphologically simple words. They found that compounds patterned with morphologically simple words rather than with adjective-noun phrases in terms of latencies, suggesting that the minimal unit of phonological encoding was the phonological word at the phrasal level, not at the lexical level and the internal prosodic structure of compounds did not affect phonological encoding for speech production. The timing of interference effects has also influenced this debate. Glaser and Düngelhoff (1984) used variable SOAs to show that semantic interference could persist even with late distractor presentation, undermining strict sequential models. Damian and Bowers (2003) added that semantic interference occurs only with word distractors, not pictures, suggesting that the competition arises at the lexical level rather than the conceptual level.

6.1.3 Tune Planning

Tune planning concerns how and when the prosodic features of speech are organised. Incremental and linear models (e.g., Levelt et al., 1999) propose a more modular, incrementally constructed process, in which phonological encoding (including word stress and syllabification) precedes the higher-level prosodic structures. In contrast, the early stage model (e.g., Keating

& Shattuck-Hufnagel, 2002) challenges this sequence by reasserting the primacy of prosody, proposing a "prosody-first" approach in which higher-level intonational and rhythmic structures are planned early and guide the insertion of segments. In short, the process begins with building a prosodic frame and then inserting segments. Lahiri (2009) investigated the planning of intonation tunes in speech production, using the PWI paradigm to determine when and how abstract intonational contours were accessed and realised. She found evidence for early retrieval of tunes, facilitating naming latencies when the intonational contour of an auditory distractor matched the intended intonation, and a later-stage effect during phonological encoding where intonation matching reduced production errors. Therefore, she proposed that intonational planning operated independently of lexical-semantic and phonological retrieval processes, supporting the hypothesis of an autonomous intonation lexicon.

6.2 Research Questions

While Lahiri (2009) focused on English, a non-tonal language, the current study extends this line of research to a tonal language, where lexical tones carry both phonological and semantic properties. Rather than analysing the prosodic features of responses, the current study uses the same experimental paradigm (PWI) as Lahiri (2009), exploring whether the abstract representation of tunes observed in English also holds in tonal systems, or whether the integration of tone and lexical meaning in such languages leads to different patterns of planning and interference.

This study addresses two key questions:

1. To what extent is a tune an independent entity in the mental lexicon?

2. Given that a single word can function as an intonational phrase, how does a tune aligned with a word interact with its phonological (segmental and tonal) and semantic information?

6.3 Methods

6.3.1 Participants

Sixty participants from Xi'an, China, took part in this study, with 20 participants in each SOA group. All participants were native speakers of GuanM, maintaining daily usage, and none reported any hearing or speaking disorders. Informed consent was obtained from all participants before the experiment began.

6.3.2 Stimuli

The experimental materials consisted of 18 degraded line drawings, with nine corresponding to T1T1 disyllabic words and nine to T2T2 disyllabic words. In addition, 18 pictures with solid lines were used as fillers. Each picture was paired with one of four types of auditory distractors: semantically related, phonologically related, semantically unrelated, or phonologically unrelated. These distractors were presented in both declarative and interrogative intonations.

For the selections of the distractors, to ensure the semantic relatedness of the distractors, 72 native speakers of GuanM participated in an online questionnaire using a Likert scale to rate the semantic closeness between the stimuli and their semantically related distractors on a scale from 1 (least close) to 7 (extremely close). The reliability of the closeness was confirmed through the use of Cronbach's Alpha, which granted a high consistency score of 0.96. Notably, the alpha value remained stable at 0.96 even when any individual stimulus was omitted from the analysis, indicating that no single stimulus significantly affected the overall reliability,

which emphasises the semantic closeness of the chosen stimuli. However, the selection of tone combinations for these semantic distractors was not controlled. As for the phonological distractors, phonologically related ones shared the initial phoneme and its tone but differed in orthographic representation. In the case of T1T1 tone combinations in experimental pictures, where a dissimilation rule changes the surface form of the first T1 Low tone to a rising tone in production, the phonologically related distractors have the same segments but not the same surface form within the first syllable. Instead, they consist of the same segments with the underlying falling tone. For T2T2 experimental pictures, it was ensured that the second tone was distinct from T2 and differed from the first tone. For unrelated distractors, they shared neither semantic nor phonological relationships with the experimental pictures, nor were the tones of these distractors controlled. The complete lists of the stimuli used in this study are presented Table 6.1 for the experimental pictures and Table 6.2 for filler pictures.

Table 6.1 Experimental picture names used in the study with their audio distractors

Picture name	**Phonological**		**Semantic**	
	Related	**Unrelated**	**Related**	**Unrelated**
医生	衣橱	牛奶	病人	地球
i31ʂəŋ31	i31tʂʰu24	niɤu24næ52	piŋ55ʐen24	ti55tɕʰiɤu24
‘doctor’	‘closet’	‘milk’	‘patient’	‘earth’
香蕉	乡党	键盘	猴子	手表
ɕiaŋ31tɕiau31	ɕiaŋ31taŋ52	tɕian55pʰan24	xɤu24tsi	ʂɤu52piau52
‘banana’	‘villager’	‘keyboard’	‘monkey’	‘watch’
西瓜	稀饭	楼梯	水果	牙刷
ɕi31kua31	ɕi31fan55	lɤu24tʰi31	fei52kuɤ52	nia24fa31
‘watermelon’	‘porridge’	‘stairs’	‘fruit’	‘toothbrush’
秋千	丘陵	快递	滑梯	咖啡
tɕʰiɤu31tɕʰian31	tɕʰiɤu31liŋ24	kʰuæ55ti55	xua24tʰi31	kʰa31fi31
‘swing’	‘hill’	‘delivery’	‘slide’	‘coffee’
蜘蛛	芝麻	领导	网子	戒指
tʂi31tʂu31	tsi31ma24	liŋ52tau52	vaŋ52tsi52	tɕiæ55tsi52
‘spider’	‘sesame’	‘leader’	‘net’	‘ring’

飞机	妃嫔	酒水	航班	报纸
fi31tɕi31	fi31pʰien24	tɕiɤu52fei52	xaŋ24pan31	pau55tsi52
‘airplane’	‘mistress’	‘beverage’	‘airline’	‘newspaper’
青蛙	清水	数学	蝌蚪	闹钟
tɕʰiŋ31ua31	tɕʰiŋ31fei52	fu55ɕyɤ24	kʰuɤ31tɤu52	nau55tʂəŋ31
‘frog’	‘clear water’	‘math’	‘tadpole’	‘alarm’
风车	丰收	字典	玩具	陕西
fəŋ31tʂʰɤ31	fəŋ31ʂɤu31	tsi55tian52	uan24tɕy55	ʂan52ɕi31
’pinwheel’	‘harvest’	‘dictionary’	‘toy’	‘Shaanxi’
乌龟	屋顶	朋友	王八	钱包
u31kuei31	u31tiŋ52	pʰəŋ24iɤu52	uaŋ24 pa31	tɕʰian24pau31
‘tortoise’	‘roof’	‘friend’	‘tortoise’	‘wallet’
学习	穴位	米饭	课本	鸡蛋
ɕyɤ24ɕi24	ɕyɛ24uei55	mi52fan55	kʰuɤ55pen52	tɕi31tan55
‘study’	‘acupoint’	‘rice’	‘textbook’	‘egg’
轮船	伦理	吸管	海洋	行李
lun24tʂʰan24	lun24li52	ɕi31kuan52	xæ52iaŋ24	ɕiŋ24li52
‘ship’	‘ethics’	‘straw’	‘sea’	‘luggage’
绵羊	棉被	板凳	动物	保安
mian24iaŋ24	mian24pi55	pan52təŋ55	tuəŋ55vu55	pau52ŋan31
‘sheep’	‘quilt’	‘stool’	‘animal’	‘security’
男人	南瓜	沙发	性别	暖气
nan24ʐen24	nan24kua31	sa31fa31	ɕiŋ55piɛ24	luan52tɕʰi55
‘man’	‘pumpkin’	‘sofa’	‘gender’	‘heating’
蓝莓	栏杆	鞭炮	果酱	游戏
lan24mei24	lan24kan52	pian31pʰau55	kuɤ52tɕiaŋ55	iɤu24ɕi55
‘blueberry’	‘railing’	‘firecracker’	‘jam’	‘game’
篮球	蓝色	艺术	比赛	天空
lan24tɕʰiɤu24	lan24sei31	i55fu55	pi52sæ55	tʰian31kʰuəŋ31
‘basketball’	‘blue’	‘art’	‘competition’	‘sky’
蝴蝶	湖水	浪漫	花朵	春晚
xu24tiɛ24	xu24fei52	laŋ55mann55	xua31tuɤ52	tʂʰen31van52
‘butterfly’	‘lake’	‘romance’	‘flower’	‘spring gala’
毛驴	矛盾	麻油	农活	钢笔
mau24ly24	mau24tuen55	ma24 iou24	nuəŋ24xuɤ24	kaŋ31pi31
‘donkey’	‘conflict’	‘sesame oil’	‘farming’	‘pen’
葡萄	菩萨	旅游	紫色	公交
pʰu24tʰau24	pʰu24sa52	ly52iɤu24	tsi52sei31	kuəŋ31tɕiau31
‘grapes’	‘servant’	‘travel’	‘purple’	‘bus’

Table 6.2 Picture names of control group used in the study with their audio distractors

Picture name	Phonological		Semantic	
	Related	Unrelated	Related	Unrelated
电脑 tian55nau52 'computer'	店员 tian55yan24 'clerk'	数学 fu55ɕyɤ24 'math'	鼠标 fu52piau31 'mouse'	牙刷 nia24fa31 'toothbrush'
椅子 i52tsi52 'chair'	已经 i52tɕiŋ31 'already'	天空 tʰian31kʰuəŋ31 'sky'	桌子 tʂuɤ31tsi 'desk'	板凳 pan52təŋ55 'stool'
奥运 ŋau55yen55 'Olympics'	澳门 ŋau55men24 'Macao'	米饭 mi52fan55 'rice'	北京 pei31tɕiŋ31 'Beijing'	鸡蛋 tɕi31tan55 'egg'
灯笼 təŋ31luəŋ24 'lantern'	登高 təŋ31kau31 'climb high'	楼梯 lɤu24tʰi31 'stairs'	元宵 yan24ɕiau31 'Lantern Festival'	手表 ʂɤu52piau52 'watch'
手机 ʂɤu52tɕi31 'phone'	首相 ʂɤu52ɕiaŋ55 'prime minister'	旅游 ly52iɤu24 'travel'	微信 vi24ɕien55 'Wechat'	牛奶 niɤu24næ52 'milk'
蘑菇 muɤ24ku31 'mushroom'	魔法 muɤ24fa31 'magic'	艺术 i55fu55 'art'	青菜 tɕʰiŋ31tsʰæ55 'cabbage'	快递 kʰuæ55ti55 'delivery'
蜡烛 la31tʂu31 'candle'	辣椒 la31tɕiau31 'pepper'	领导 liŋ52tau52 'leader'	停电 tʰiŋ24tian55 'black-out'	公交 kuəŋ31tɕiau31 'bus'
拖鞋 tʰuɤ31xæ24 'slippers'	托盘 tʰuɤ31pʰan24 'tray'	闹钟 nau55tʂəŋ31 'alarm'	室内 ʂi31luei24 'indoors'	咖啡 kʰa31fi31 'coffee'
雨伞 y52san52 'umbrella'	语言 y52nian24 'language'	酒水 tɕiɤu52fei52 'beverage'	折叠 tʂɤ31tiɛ24 'folding'	陕西 ʂan52ɕi31 'Shaanxi'
月亮 yɛ31liaŋ55 'moon'	阅读 yɛ31tu24 'read'	字典 tsi55tian52 'dictionary'	星星 ɕiŋ31ɕiŋ31 'star'	朋友 pʰəŋ24iɤu52 'friend'
相机 ɕiaŋ55tɕi31 'camera'	项目 ɕiaŋ55mu31 'programme'	钱包 tɕʰian24pau31 'wallet'	照片 tʂau55pʰian55 'photo'	吸管 ɕi31kuan52 'straw'
汽车 tɕʰi55tʂʰɤ31 'car'	契约 tɕʰi55nyɤ31 'contract'	春晚 tʂʰen31van52 'spring gala'	上班 ʂaŋ55pan31 'go to work'	暖气 luan52tɕʰi55 'heating'

剪刀	简单	游戏	工具	地球
tɕian52tau31	tɕian52tan31	iɤu24ɕi55	kuəŋ31tɕy55	ti55tɕʰiɤu24
‘scissors’	‘simplicity’	‘game’	‘tool’	‘earth’
耳机	洱海	保安	听歌	键盘
ɚ52tɕi31	ɚ52xæ52	pau52ŋan31	tʰiŋ31kɤ31	tɕian55pʰan24
‘earphones’	‘Erhai’	‘security’	‘listen to music’	‘keyboard’
老鼠	姥爷	浪漫	耗子	麻油
lau52fu52	lau52iɛ24	laŋ55mann55	xau55tsi52	ma24 iou24
‘rat’	‘grandpa’	‘romance’	‘rat’	‘sesame oil’
骆驼	落叶	报纸	沙漠	行李
luɤ31tʰuɤ24	luɤ31iɛ31	pau55tsi52	sa31mei31	ɕiŋ24li52
‘camel’	‘fallen leaves’	‘newspaper’	‘desert’	‘luggage’
蛋糕	淡水	鞭炮	生日	沙发
tan55kau31	tan55fei52	pian31pʰau55	ʂəŋ31i31	sa31fa31
‘cake’	‘fresh water’	‘firecracker’	‘birthday’	‘sofa’
火车	伙伴	钢笔	旅游	戒指
xuɤ52tʂʰɤ31	xuɤ52pan55	kaŋ31pi31	ly52iɤu24	tɕiæ55tsi52
‘train’	‘partner’	‘pen’	‘travel’	‘ring’

The auditory distractors were recorded by a 30-year-old female native speaker of GuanM from the same region where the study was conducted, at a frequency of 44.1 kHz, using the software Audacity. The experiment had 288 trials in total (18 degraded line drawings and 18 solid line pictures, each paired with two distractor intonations, two distractor types, and two distractor relatedness conditions), which were divided into eight blocks, with no repetition of pictures within the same block. Distractors were presented at three different SOAs: -300 ms (before picture onset), 0 ms (simultaneously with picture onset), and 300 ms (after picture onset). Participants were randomly assigned to one of the SOA groups.

Two examples of experimental items including T1T1 [ɕi31kua31] ‘西瓜, watermelon’ and T2T2 [lan24 tɕʰiɤu24] ‘篮球, basketball’, along with their respective distractor conditions are presented in Figure 6.1 and Figure 6.2. All experimental pictures were presented with degraded

lines to elicit a question intonation, while distractors were presented at the three different SOAs relative to picture onset.

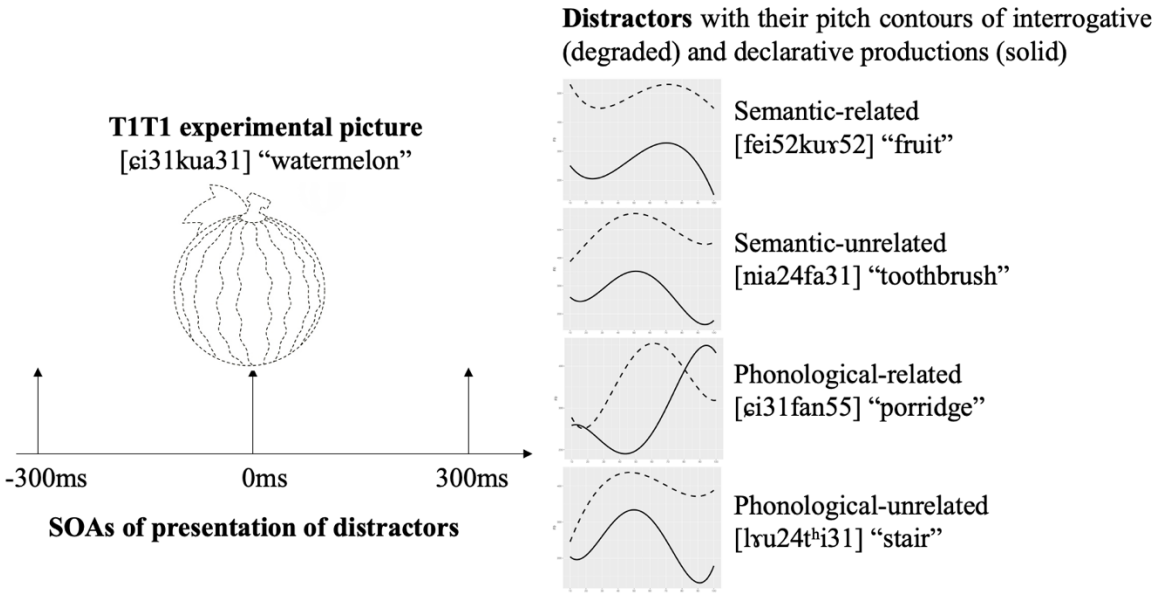


Figure 6.1 Example of one T1T1 experimental picture with its four types of distractors with both intonations

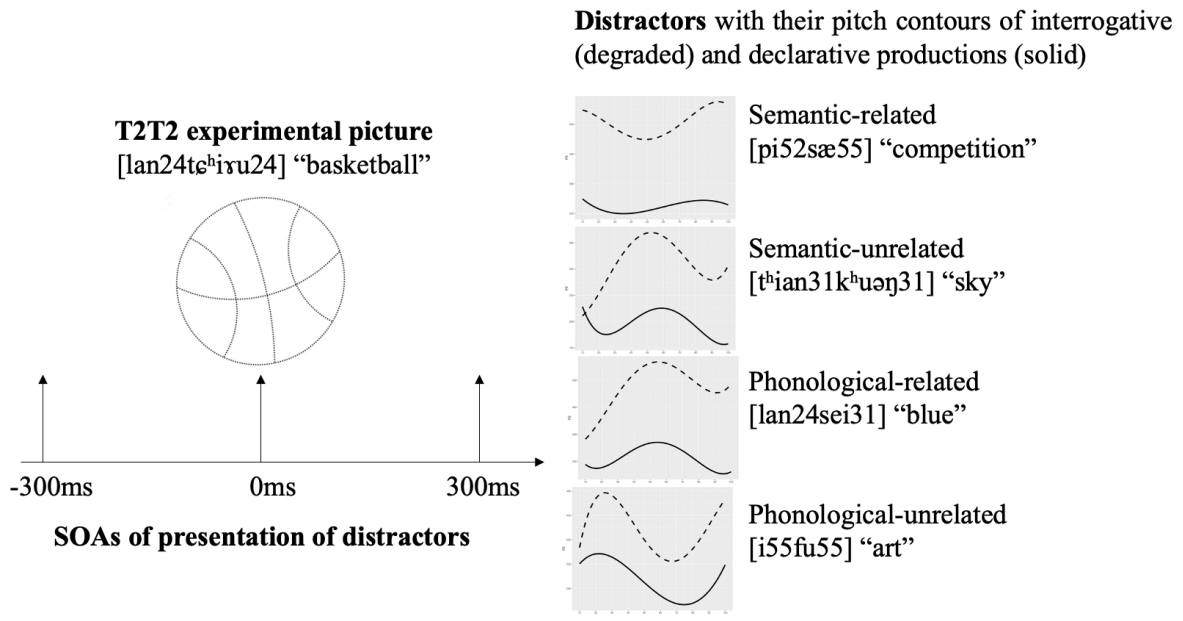


Figure 6.2 Example of one T2T2 experimental picture with its four types of distractors with both intonations

6.3.3 Procedure

The experiment, programmed using PsychoPy (Peirce et al., 2019), was conducted in a quiet room in a local community centre. Each trial began with a 500 ms fixation cross, followed by a picture displayed for 1500 ms. During this time, distractor audio was played through headphones at one of three experimental SOAs (-300 ms, 0 ms, or 300 ms). Participants had 3000 ms to respond after each picture's onset. They were instructed to name the pictures with degraded lines using an interrogative intonation and those with solid lines (fillers) using a declarative intonation, as quickly and accurately as possible while ignoring the auditory distractors. Before the main experiment, participants familiarised themselves with all the pictures and had 10 practice trials with different images. Responses were recorded using the built-in microphone of a Lenovo laptop, starting at the onset of each picture display. A 500 ms fixation cross (+) was inserted between trials. The experimental procedure for an example trial is shown in Figure 6.3. The entire experiment took approximately one hour to complete. Participants were allowed to take breaks between eight blocks (if desired) and were compensated for their time.

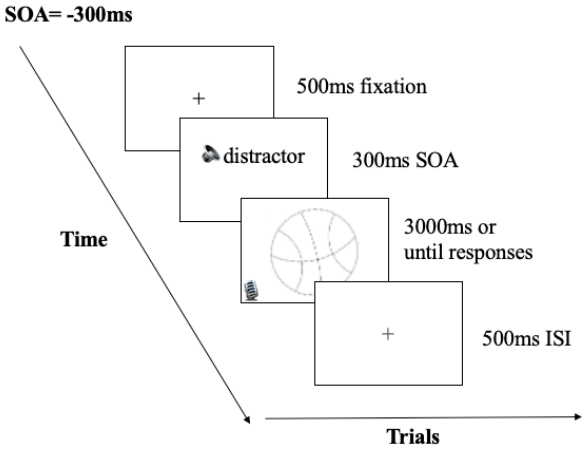


Figure 6.3 Illustration of the temporal sequence of experimental pictures at SOA -300 ms

6.3.4 Statistical Analyses

The audio responses were recorded and saved as individual files, capturing the period from the onset of the pictures to the presentation to the next trial. Reaction times (RTs) were measured from the onset of the picture display to the onset of the oral response given by the participant, using Chronset (Roux et al., 2017), a tool developed for automatic speech onset detection. These RT measurements were then manually corrected by a trained phonetician in Praat (Paul Boersma & Weenink, 2023). Responses were only coded as correct if both the correct picture name and the correct tune were produced.

Statistical analyses of the RTs were conducted using the linear mixed-effects models fitted in R (R Core Team, 2024) using the lmerTest package (Kuznetsova et al., 2017). The dependent variable was RT, with fixed effects that varied according to the level of analysis. The test began with broad analyses examining the overall effect of Distractor Intonation (interrogative vs. declarative) on picture naming, then progressively narrowed the focus to examine tone-specific effects (T1T1 vs. T2T2), SOA-specific effects (-300 ms, 0 ms, 300 ms), and finally the interactions between Distractor Type (phonological vs. semantic), and Relatedness (related vs.

unrelated) separately for both Distractor Intonation across each SOA within each Tone. All models included random effects for Participant and Item. Post-hoc comparisons of significant interactions were conducted using the emmeans package in R (Lenth, 2024) for pairwise comparisons with Tukey's correction. The RT model for the interrogative tune, fitted separately for each SOA within each Tone, was specified in R as follows:

Model_RT <- lmer(RT ~ Distractor_Type * Distractor_Relatedness + (1 | Participant) + (1 | Item), data = Interrogative_Tune)

For accuracy data, the generalised mixed-effects models with the same structure as the RT analyses were applied. All effects were considered statistically significant at $p < .05$, with marginal significance noted for *p*-values between .05 and .1.

6.4 Results

Due to a technical issue during data collection, one participant's block of 36 trials at SOA 300 ms was not saved, reducing the dataset to 17,244 data points across three SOAs. Before proceeding with the statistical analyses, outliers were removed by excluding RTs that were more than ±3 standard deviations from each participant's mean RT within each block, which led to the exclusion of 1.54% of the data (265 trials). This resulted in 16,979 valid responses, of which 8,497 were responses to filler items and 8,482 were to experimental pictures, with 4,225 from T1T1 and 4,257 from T2T2.

6.4.1 Fillers

The analysis began with an examination of accuracy and RTs for the filler trials in which participants were instructed to name pictures using a declarative tune, with audio distractors played in both interrogative and declarative tunes. This process aimed to create a baseline for

the more complicated experimental trials related to tone. Since the tone combinations for the filler images were not controlled, only the accuracy and RTs concerning Distractor Intonation across three SOAs were examined.

6.4.1.1 Accuracy

Overall, accuracy remained consistently high, exceeding 96% across all three SOAs. Statistical analysis revealed no significant main effect of Distractor Intonation on accuracy across the SOAs (all *p*s > .05). Detailed accuracy rates by Distractor Intonation at each SOA are presented in Table 6.3.

Table 6.3 Accuracy of fillers by the Distractor Intonation at three SOAs

SOA	Distractor Intonation	N	Mean (SD)
-300 ms	Declarative	1416	0.96 (0.20)
	Interrogative	1419	0.96 (0.19)
0 ms	Declarative	1420	0.98 (0.15)
	Interrogative	1421	0.98 (0.15)
300 ms	Declarative	1411	0.97 (0.17)
	Interrogative	1410	0.97 (0.18)

6.4.1.2 RTs

No significant main effect of Distractor Intonation on RTs was found ($\chi^2(1) = .08$, $p = .778$), meaning that RTs did not significantly differ based on whether the distractor tunes were declarative (mean = 920.37 ms) or interrogative (mean = 919.51 ms) when naming pictures in a declarative tune. This consistency in RTs, regardless of Distractor Intonation, held across all SOAs (all *p*s > .05). Refer to Table 6.4 for the RTs by Distractor Intonation across the three SOAs, suggesting that the declarative tune may serve as a default form. When producing a declarative tune, RTs were not significantly impacted by whether the distractors had a declarative or interrogative tune.

Table 6.4 RTs of fillers by the Distractor Intonation at three SOAs

SOA	Distractor Intonation	N	Mean (SD)
-300 ms	Declarative	1359	894.87 (205.37)
	Interrogative	1367	891.82 (211.65)
0 ms	Declarative	1389	944.54 (211.70)
	Interrogative	1389	944.13 (229.73)
300 ms	Declarative	1370	921.16 (268.88)
	Interrogative	1364	922.18 (264.08)

6.4.2 Experimental Pictures

6.4.2.1 Accuracy

The main effect of Distractor Intonation was not significant ($\chi^2(1) = 0.04$, $p = .838$). Accuracy was high across both tunes, with 96.4% for declarative tune distractors and 96.5% for interrogative tune distractors. When examined across the three SOAs for both tones, the main effect of Distractor Intonation on accuracy was consistently non-significant (all ps > .05). This result indicates that both declarative and interrogative tune distractors produced similarly high accuracy. See Table 5.4 for detailed accuracy rates by Distractor Intonation and tone across the three SOAs.

Table 6.5 Accuracy of experimental pictures by the Distractor Intonation at three SOAs for T1T1 and T2T2

SOA	Distractor Intonation	T1T1		T2T2	
		N	Mean (SD)	N	Mean (SD)
-300 ms	Declarative	708	0.97 (0.17)	713	0.95 (0.23)
	Interrogative	698	0.97 (0.18)	708	0.95 (0.21)
0 ms	Declarative	704	0.98 (0.14)	717	0.96 (0.19)
	Interrogative	711	0.98 (0.15)	707	0.96 (0.18)
300 ms	Declarative	701	0.97 (0.16)	706	0.95 (0.21)
	Interrogative	703	0.97 (0.17)	706	0.96 (0.20)

The analysis of accuracy was further divided between T1T1 and T2T2 to examine the effects of Distractor Type and Relatedness on each Distractor Intonation separated by tones, given that T1 (low tone) and T2 (rising tone) have distinct tone contour shapes. For T1T1 (see Table 6.6), the statistical results indicated that the main effects of Distractor Type, Relatedness, and their interaction were not significant across both tunes and all three SOAs (all *p*s > .05). Accuracy remained consistently high, exceeding 95% across all conditions.

Table 6.6 Accuracy for T1T1 across three SOAs by Distractor Intonation and Distractor Type

Distractor Intonation	Distractor Type	Relatedness	SOA -300 ms	SOA 0 ms	SOA 300 ms
Declarative	Phonological	Related	0.96 (0.19)	0.99 (0.11)	0.98 (0.15)
		Unrelated	0.97 (0.17)	0.98 (0.15)	0.99 (0.11)
	Semantic	Related	0.97 (0.17)	0.97 (0.17)	0.97 (0.18)
		Unrelated	0.98 (0.15)	0.98 (0.13)	0.97 (0.18)
Interrogative	Phonological	Related	0.98 (0.13)	0.98 (0.15)	0.97 (0.17)
		Unrelated	0.97 (0.18)	0.98 (0.15)	0.97 (0.18)
	Semantic	Related	0.95 (0.22)	0.96 (0.19)	0.96 (0.20)
		Unrelated	0.97 (0.17)	0.99 (0.11)	0.98 (0.15)

For T2T2 (see Table 6.7), the main effects of Distractor Type, Relatedness, and their interaction were not significant across both tunes and all three SOAs (all ps > .05), with the exception of the declarative tune at SOA 300 ms. At this SOA, the main effect of Distractor Type was significant ($\chi^2(2) = 6.01$, $p = .050$), and the interaction between Distractor Type and Relatedness was also significant ($\chi^2(1) = 5.00$, $p = .025$). In addition, the main effect of Relatedness showed a trend towards significance ($\chi^2(2) = 5.37$, $p = .068$). Within semantic distractors, related distractors tended to result in lower accuracy than unrelated distractors ($\beta = -1.17$, $z = -1.87$, $p = .061$), suggesting a possible trend towards a semantic interference effect. When comparing

Distractor Types within each level of Relatedness, phonological-related distractors had significantly higher accuracy than semantic-related distractors (β = 1.40, z = 2.15, p = .032).

Meanwhile, at SOA 300 ms with the interrogative tune, no main effects reached statistical significance, although the main effect of Relatedness showed a trend towards significance ($\chi^2(2) = 5.09$, p = .078), indicating a potential influence of Relatedness on accuracy. Pairwise comparisons for phonological distractors revealed a significant difference between related and unrelated items, with phonological-related distractors showing higher accuracy (β = 1.46, z = 2.04, p = .041), suggesting a phonological facilitation effect on accuracy at SOA 300 ms.

Table 6.7 Accuracy for T2T2 across three SOAs by Distractor Intonation and Distractor Type

Distractor Intonation	Distractor Type	Relatedness	SOA -300 ms	SOA 0 ms	SOA 300 ms
Declarative	Phonological	Related	0.96 (0.19)	0.98 (0.15)	0.97 (0.17)
		Unrelated	0.94 (0.23)	0.95 (0.22)	0.95 (0.22)
	Semantic	Related	0.92 (0.28)	0.95 (0.22)	0.93 (0.26)
		Unrelated	0.96 (0.19)	0.97 (0.18)	0.97 (0.18)
Interrogative	Phonological	Related	0.96 (0.19)	0.98 (0.13)	0.98 (0.15)
		Unrelated	0.96 (0.20)	0.97 (0.18)	0.94 (0.24)
	Semantic	Related	0.93 (0.25)	0.97 (0.17)	0.97 (0.18)
		Unrelated	0.96 (0.20)	0.94 (0.24)	0.95 (0.21)

6.4.2.2 RTs

6.4.2.2.1 Tune Effect

Before investigating the interaction between tone and tune in tune planning, the overall effect of Tune was examined by aggregating all distractors that differed only in intonation. The effect for Distractor Intonation was significant ($\chi^2(1) = 11$, $p < .001$). RTs for distractors with an interrogative tune (mean = 946.61 ms) were significantly faster than those with a declarative

tune (mean = 958.99 ms), suggesting a facilitative effect of interrogative intonation on cognitive processing.

Divided into different tone groups, the effect of Distractor Intonation was also significant within T1T1 ($\chi^2(1) = 5.46$, $p = .019$) and T2T2 ($\chi^2(1) = 5.74$, $p = .017$). In T1T1, the mean RT in the presence of a declarative tune distractor was 969.21 ms, while the mean RT with an interrogative tune distractor was 956.77 ms. The interrogative tune significantly reduced RTs by 12.44 ms. Similarly, in T2T2, RTs for distractors with an interrogative intonation (mean = 936.36 ms) were significantly shorter by 12.29 ms than ones with a declarative intonation (mean = 948.65 ms), confirming the pattern of faster responses with interrogative intonations across both groups during the production of the interrogative tune.

Breaking down the data across the three SOAs, RTs were consistently faster for interrogative tune distractors compared to declarative ones (see Table 6.8), suggesting a facilitative effect when the intonational tune of the distractor matched the intended question tune production. This effect reached statistical significance at SOA 0 ms ($\chi^2(1) = 8.75$, $p = .003$), where interrogative distractors led to a reduction of 19.91 ms in RTs. At SOA 300 ms, the effect was also significant ($\chi^2(1) = 4.34$, $p = .037$), with RTs 14.95 ms shorter for interrogative distractors than declarative distractors. However, at SOA -300 ms, there was no evidence of a significant effect ($\chi^2(1) = 0.46$, $p = .500$).

Table 6.8 RTs of experimental pictures by the Distractor Intonation at three SOAs

SOA	Distractor Intonation	N	Mean (SD)
-300 ms	Declarative	1322	930.06 (214.26)
	Interrogative	1312	925.55 (217.85)

0 ms	Declarative	1340	982.94 (218.45)
	Interrogative	1338	963.59 (213.93)
300 ms	Declarative	1316	963.66 (283.69)
	Interrogative	1321	950.33 (269.57)

Furthermore, the faster RTs with a distractor in an interrogative tune was observed across both T1T1 and T2T2 tone groups, though the degree of facilitation effect varied by Tone and SOA. At SOA -300 ms, no significant effects were found in either group (T1T1: $\chi^2(1) = 0.74$, $p = .389$; T2T2: $\chi^2(1) = 0.02$, $p = .884$). At SOA 0 ms, for T2T2, a significant main effect of Distractor Intonation was observed ($\chi^2(1) = 6.05$, $p = .014$), with interrogative distractors reducing RTs by 22.63 ms, indicating clear facilitation. For T1T1, the effect at SOA 0 ms only approached significance ($\chi^2(1) = 3.06$, $p = .080$), with a slightly smaller RT reduction of 16.97 ms, suggesting a possible but weaker facilitative effect. At SOA 300 ms, neither group showed a statistically significant effect of Distractor Intonation on RTs, although numerical trends were consistent with the earlier pattern (T1T1: $\chi^2(1) = 2.07$, $p = .151$; T2T2: $\chi^2(1) = 2.28$, $p = .131$). See Table 6.9 for the mean RTs by Distractor Intonation for both T1T1 and T2T2.

Table 6.9 RTs of experimental pictures by Distractor Intonation at three SOAs for T1T1 and T2T2

SOA	Distractor Intonation	T1T1		T2T2	
		N	Mean (SD)	N	Mean (SD)
-300 ms	Declarative	648	944.92 (224.13)	674	914.92 (202.75)
	Interrogative	637	937.41 (221.53)	675	913.68 (213.61)
0 ms	Declarative	651	995.96 (225.52)	689	969.91 (210.49)
	Interrogative	656	978.18 (224.90)	682	948.74 (201.24)
300 ms	Declarative	643	966.62 (288.11)	673	960.66 (279.31)
	Interrogative	643	954.14 (274.56)	678	946.51 (264.60)

6.4.2.2.2 Tone Interaction (T1T1)

Figure 6.4 presents the RTs of T1T1 by Distractor Intonation, Distractor Type, and Relatedness. At SOA -300 ms, significant effects were found for the declarative tune across all fixed effects. Specifically, both Distractor Type ($\chi^2(2) = 15.95$, $p < .001$) and Relatedness ($\chi^2(2) = 34.79$, $p < .001$) showed significant main effects on RTs, along with a significant interaction between Distractor Type and Relatedness ($\chi^2(1) = 8.27$, $p = .004$). For phonological distractors, related items had significantly longer RTs than unrelated items, indicating a strong phonological inhibition effect ($\beta = 102.87$, $t(656.06) = 5.73$, $p < .001$). Participants tended to respond slower to related than unrelated semantic distractors ($\beta = 29.72$, $t(656.11) = 1.65$, $p = .099$), suggesting a weaker semantic inhibition effect. In addition, RTs for phonological-related distractors were significantly slower than for semantic-related ones ($\beta = 71.98$, $t(656.07) = 4.01$, $p < .001$). In the interrogative tune condition, neither the main effect of Distractor Type ($\chi^2(2) = 0.76$, $p = .683$) nor the interaction between Distractor Type and Relatedness ($\chi^2(1) = 0.14$, $p = .707$) was significant. There was a marginally significant main effect of Relatedness ($\chi^2(2) = 4.97$, $p = .083$). Further post-hoc testing showed that phonological-related distractors tended to lead to slower RTs compared to phonological-unrelated distractors ($\beta = 33.96$, $t(644.22) = 1.82$, $p = .069$), suggesting a trend towards phonological interference in the interrogative tune condition.

In the case of the declarative tune at SOA 0 ms, there were no statistically significant main effects for either Distractor Type ($\chi^2(2) = 5.64$, $p = .060$) or Relatedness ($\chi^2(2) = 2.88$, $p = .237$), nor was their interaction significant ($\chi^2(1) = 0.69$, $p = .407$). In the interrogative tune condition, the main effect of Distractor Type was significant ($\chi^2(2) = 8.79$, $p = .012$), while the main effect of Relatedness was marginally significant ($\chi^2(2) = 5.53$, $p = .063$). The interaction between Distractor Type and Relatedness was significant ($\chi^2(1) = 4.99$, $p = .026$). To be specific, RTs were significantly longer for semantic ones rather than unrelated distractors ($\beta = 40.70$,

$t(663.11) = 2.10, p = .036$), suggesting a semantic inhibition effect. When comparing Distractor Types within each level of Relatedness, semantic-related distractors were found to have significantly longer RTs than phonological-related distractors ($\beta = -57.39$, $t(663.14) = -2.96$, $p = .003$), suggesting that semantic-related distractors cause greater interference than phonological-related ones.

There were significant main effects of both Distractor Type ($\chi^2(2) = 14.55$, $p < .001$) and Relatedness ($\chi^2(2) = 19.45$, $p < .001$), for the declarative tune at SOA 300 ms, as well as a significant interaction between Distractor Type and Relatedness ($\chi^2(1) = 9.78$, $p = .002$). Participants took significantly longer to respond to related phonological distractors than to unrelated distractors ($\beta = 89.99$, $t(652.11) = 4.43$, $p < .001$), showing a phonological inhibition effect. When comparing phonological and semantic distractors, pictures with phonological-related distractors elicited significantly longer RTs than semantic-related distractors ($\beta = 77.35$, $t(652.09) = 3.77$, $p < .001$), suggesting a phonological-specific delay in response. No significant main effect for Distractor Type ($\chi^2(2) = 5.30$, $p = .071$) or Relatedness ($\chi^2(2) = 3.34$, $p = .189$) was found for the interrogative tune at SOA 300 ms, nor was their interaction significant ($\chi^2(1) = 2.49$, $p = .115$) Related items tended to have longer response times than unrelated items for semantic distractors; however, this difference was not statistically significant ($\beta = 35.85$, $t(650.18) = 1.77$, $p = .078$), suggesting a potentially weak semantic interference effect, and RTs for semantic-related distractors were significantly slower than for phonological-related distractors ($\beta = -46.48$, $t(650.15) = -2.30$, $p = .022$).

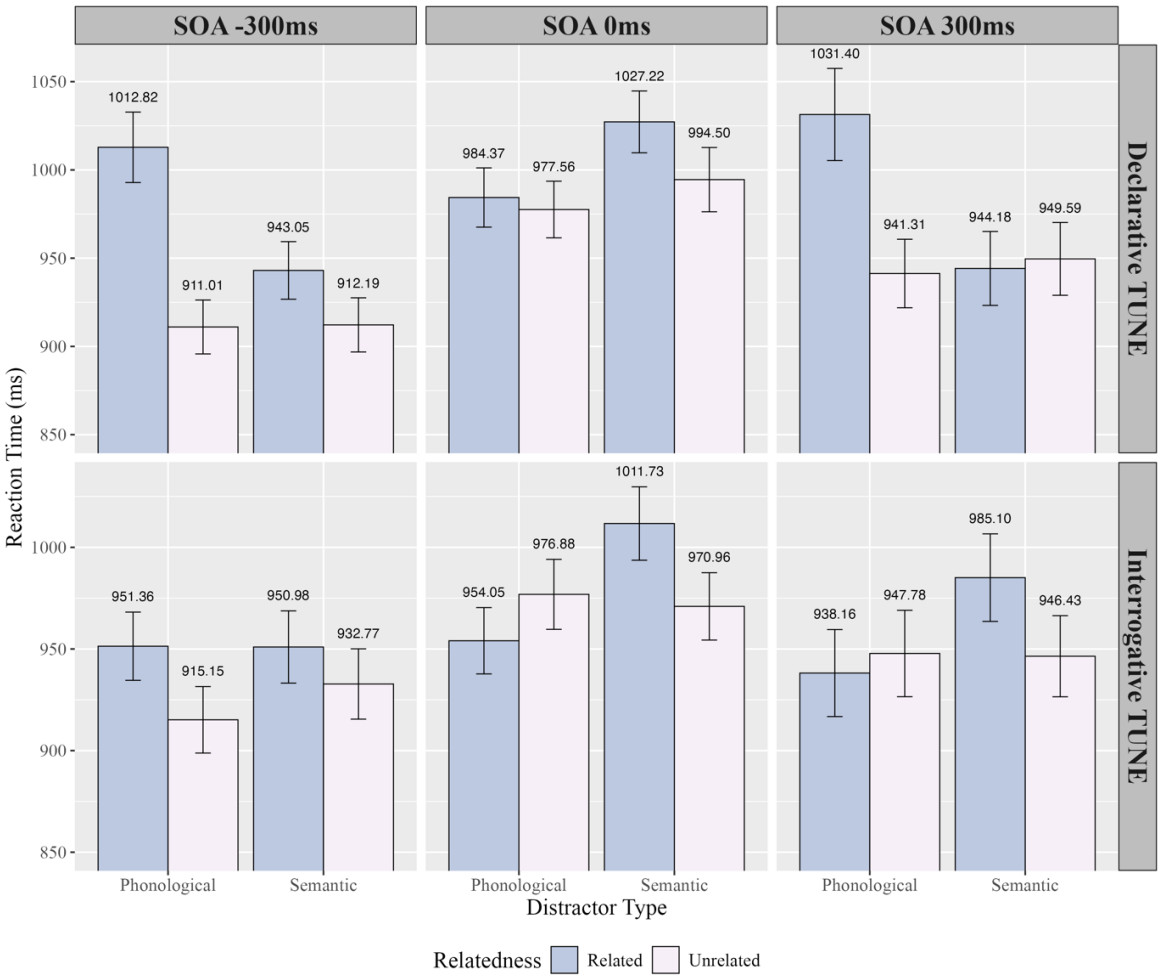


Figure 6.4 RTs for T1T1 across SOAs by Distractor Intonation and Distractor Type

6.4.2.2.3 Tone Interaction (T2T2)

For T2T2, (see Figure 6.5), with the declarative tune at SOA -300 ms, there were no significant effects observed for Distractor Type ($\chi^2(2) = 1.73, p = .421$), Relatedness ($\chi^2(2) = 2.10, p = .350$) or for the interaction between Distractor Type and Relatedness ($\chi^2(1) = 0.97, p = .325$), suggesting that neither Relatedness nor the combined effect of Distractor Type and Relatedness influenced RTs when distractors had a declarative tune. For the interrogative tune, significant main effects were found for both Distractor Type ($\chi^2(2) = 10.4, p = .006$) and Relatedness ($\chi^2(2) = 7.31, p = .026$), as well as for their interaction ($\chi^2(1) = 5.86, p = .016$). Post-hoc pairwise comparisons further indicated that, for semantic distractors, participants responded

significantly slower to related items compared to unrelated items (β = 47.95, t(644.19) = 2.57, p = .010), indicating a semantic interference effect. Furthermore, participants named pictures in the interrogative tune with phonological-related distractors significantly faster compared to semantic-related distractors (β = -59.92, t(644.20) = -3.22, p = .001).

For the declarative tune at SOA 0 ms, the same as observed at SOA -300 ms, neither the main effects nor the interaction between Distractor Type and Relatedness significantly influenced response times (Distractor Type: $\chi^2(2) = 0.22$, p = .896; Relatedness: $\chi^2(2) = 1.08$, p = .582; Distractor Type × Relatedness: $\chi^2(1) = 0.18$, p = .674). However, for the interrogative tune, significant main effects and interactions were found in RTs with respect to both Distractor Type and Relatedness and their interaction (Distractor Type: $\chi^2(2) = 11.29$, p = .004; Relatedness: $\chi^2(2) = 9.04$, p = .011; Distractor Type × Relatedness: $\chi^2(1) = 7.68$, p = .006), indicating that the influence of Relatedness on response times was moderated by the Distractor Type. To be specific, RTs with phonologically related distractors were significantly shorter than phonologically unrelated ones (β = -48.55, t(651.11) = -2.78, p = .006), implying a phonological facilitation effect. However, for semantic distractors, no significant difference in RTs was found between related and unrelated distractors (β = 20.41, t(651.20) = 1.15, p = .250). Further comparisons between related phonological and semantic distractors showed that phonological-related distractors led to significantly shorter RTs than semantic-related distractors (β = -57.63, t(651.12) = -3.31, p = .001), emphasising the phonological facilitative effect in enhancing response speed relative to semantic distractors.

At SOA 300 ms for the declarative tune, both main effects of Distractor Type ($\chi^2(2) = 5.77$, p = .056) and Relatedness ($\chi^2(2) = 5.76$, p = .056) showed a trend towards significance. However, the interaction effect between Distractor Type and Relatedness was significant ($\chi^2(1) = 5.48$, p

= .019). RTs for phonological distractors did not significantly differ between related and unrelated distractors (β = -25.92, t(642.17) = -1.27, p = .206), while for semantic distractors, related distractors had significantly longer RTs than unrelated ones (β = 42.30, t(642.36) = 2.04, p = .042), suggesting a semantic inhibition effect. For the interrogative tune, similar to the observations at SOA 0 ms, all main effects and interactions for RTs related to Distractor Type and Relatedness were significant (Distractor Type: $\chi^2(2) = 30.58$, $p < .001$; Relatedness: $\chi^2(2) = 16.04$, $p < .001$; Distractor Type × Relatedness: $\chi^2(1) = 14.20$, $p < .001$), indicating that the effect of Relatedness on RTs depended on the Distractor Type. Phonological distractors showed a trend where related distractors led to faster RTs than unrelated ones, though this was not statistically significant (β = -31.72, t(646.08) = -1.69, p = .091), suggesting a potential weak phonological facilitation effect, while semantic-related distractors had significantly longer RTs than unrelated items (β = 68.87, t(646.08) = 3.65, $p < .001$), indicating semantic interference. Lastly, phonological-related distractors had significantly faster RTs compared to semantic-related ones (β = -104.60, t(646.09) = -5.58, $p < .001$).

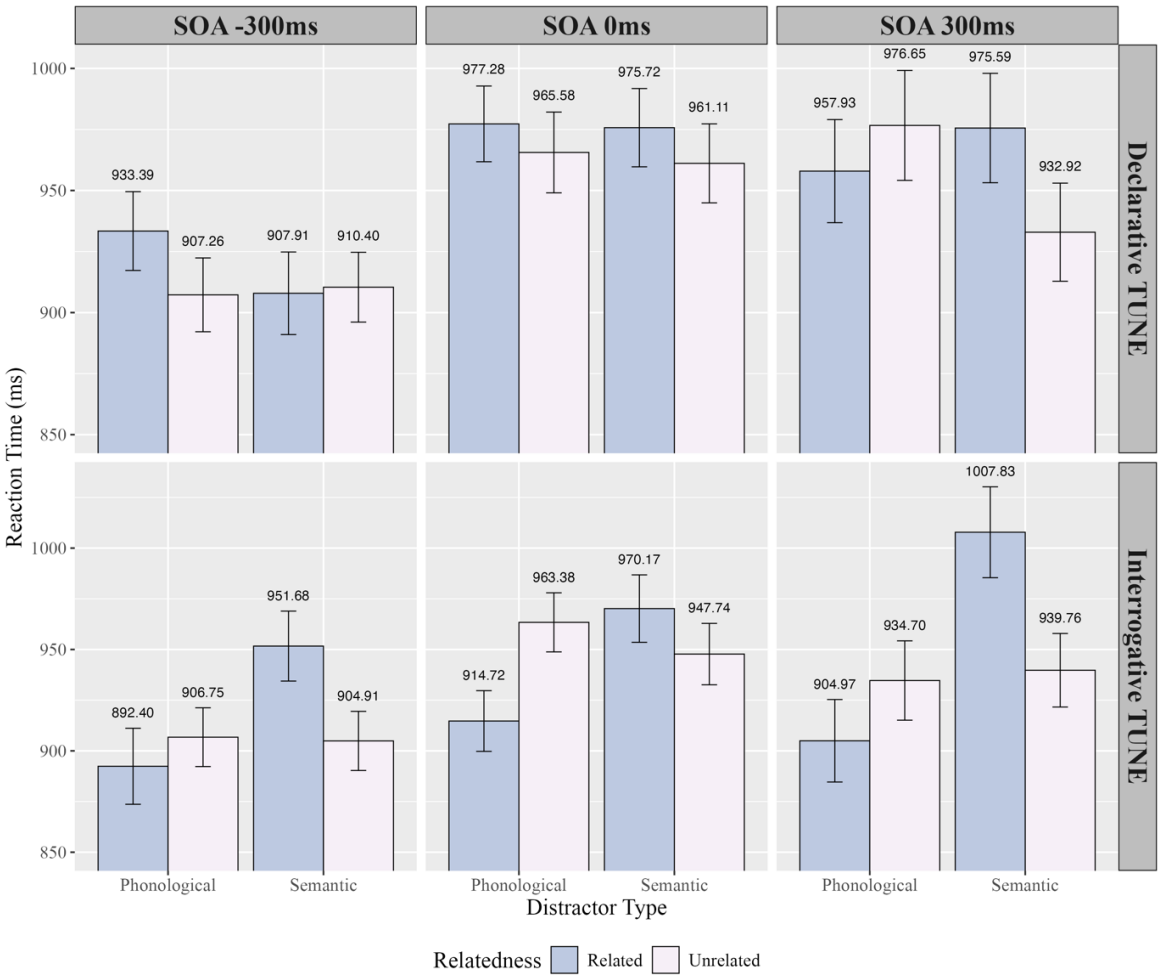


Figure 6.5 RTs for T2T2 across SOAs by Distractor Intonation and Distractor Type

6.5 Discussion

6.5.1 Tune Effects

The high accuracy of the experimental pictures and filler items validated the reliability of the study, indicating that the participants understood the tasks and were able to respond accordingly. Based on this, the observation that declarative tunes yielded no significant differences in RTs across distractor conditions suggests that the declarative intonation functions as a default or unmarked prosodic form in GuanM. The declarative tune may be pre-specified or internally maintained, even in the presence of competing prosodic cues within the speech planning architecture that require minimal additional processing.

When participants named pictures using an interrogative tune, the RTs were significantly shorter when the distractor also had an interrogative tune, compared to when the distractor had a declarative tune. The planning of the interrogative tune is influenced by the congruence between the tune of the distractor and the target, with the matching interrogative tune providing a facilitation effect, even across different SOAs and tones. The consistent facilitation effect observed when interrogative tunes in the distractor matched the required intonation of the target proves that tune planning begins early in the production process, likely at the conceptual message construction stage. Speakers retrieve and commit to an intended tune before lexical access is complete, supporting psycholinguistic models such as Levelt's (1989) that include a prosody generator but extends them by showing that tune selection may precede lemma retrieval, rather than follow it. Importantly, this effect was observed across all SOAs, including early onset times, indicating that tune congruence operates during early-stage planning rather than developing as a late-stage phonetic adjustment. Such earlier tune effect confirmed what Lahiri's (2009) discovery about the early planning of the tune.

This study aimed to clarify the role of tone in speech planning, particularly whether it interacts with semantic, phonological, and intonational information. Therefore, the analysis was separated by tone group: T1T1, which features a rising-falling surface contour due to tone dissimilation, and T2T2, which exhibits a rising-rising contour, reflecting a different pitch configuration. Based on statistically significant findings in the previous section, Table 6.10 summarises semantic and phonological effects based on RT results divided by Tone, SOA and the Distractor Intonation. In the table, effects marked as "weak" indicate *p*-values greater than .05 but less than .10, suggesting a tendency towards significance.

Table 6.10 Summary of semantic and phonological effects across three SOAs

T1T1	**-300 ms**	**0 ms**	**300 ms**
Tune clashes (declarative tune)	Phonological inhibition & Weak semantic inhibition		Phonological inhibition
Tune matches (interrogative tune)	Weak phonological inhibition	Semantic inhibition	Weak semantic inhibition
T2T2	**-300 ms**	**0 ms**	**300 ms**
Tune clashes (declarative tune)			Semantic inhibition
Tune matches (interrogative tune)	Semantic inhibition	Phonological facilitation	Weak phonological facilitation & Semantic inhibition

The table below (Table 6.11) presents the findings more clearly by using symbols to represent the types of effects observed in RTs across the T1T1 and T2T2 tone groups. ✗ indicates inhibition (i.e., slower RTs), ✓ represents facilitation (i.e., faster RTs), and - denotes a weak effect (*p*-value between .05 and .01, suggesting a tendency towards significance). P stands for phonological, and S stands for semantic, indicating the type of distractors involved.

Table 6.11 Visual summary of semantic and phonological effects across three SOAs

T1T1	**-300 ms**	**0 ms**	**300 ms**
Tune clashes (declarative tune)	P ✗ S ✗-		P ✗
Tune matches (interrogative tune)	P ✗-	S ✗	S ✗-
T2T2	**-300 ms**	**0 ms**	**300 ms**
Tune clashes (declarative tune)			S ✗
Tune matches (interrogative tune)	S ✗	P ✓	P ✓- S ✗

6.5.2 Tone Effects

In the following section, we present a stage-by-stage analysis of the processes occurring at each SOA. For the target pictures with T1T1, when declarative tune distractors precede the target at SOA -300 ms, we observe both phonological inhibition, which is likely caused by the direct competition between the distractor's falling pitch contour and the target's obligatory rising surface form for the first syllable. The declarative tune's global low register creates further misalignment with the high register required for interrogative production. The observed weaker semantic effect suggests that while conceptual processing occurs simultaneously, the predominant cognitive burden involves resolving the multi-level prosodic conflict.

In tune match conditions at this early stage, the observation of weak phonological inhibition rather than facilitation is found. Despite both distractor and target sharing interrogative tunes, the underlying tonal representations remain influential, suggesting that even when broader intonational contours align, conflicts at the tone level-particularly between underlying representations and required surface forms-can still disrupt processing, meaning that speakers simultaneously manage multiple layers of prosodic information during early planning stages.

At SOA 0 ms, when distractor and target appear simultaneously, tune clash conditions for T1T1 show no significant effects, suggesting that participants have established a commitment to their prosodic plan for interrogative production, creating temporary resistance to competing declarative tunes. The timing coincides with lexical selection processes, where attentional resources may shift from prosodic planning to lexical retrieval.

In contrast, tune match conditions at this timing produce clear semantic inhibition. The shift from phonological to semantic effects indicates that when prosodic demands are satisfied

through matching interrogative tunes, processing resources become available for lexical-semantic operations. The matching tune appears to enhance activation of the distractor's semantic content, creating stronger competition that requires active suppression during lexical selection.

The reemergence of phonological inhibition for T1T1 when distractors have declarative tunes at SOA 300 ms suggests that articulation planning remains vulnerable to tone conflicts even after lexical selection. When the declarative distractor appears during these late planning stages, its falling contour interferes with the phonetic adjustments required for T1T1's complex rising-falling pattern, particularly within an interrogative framework requiring raised pitch.

For tune match conditions at this late stage, the weak semantic inhibition indicates that semantic competition naturally weakens over time but persists through monitoring and articulation planning stages. The reduction from full to weak inhibition (compared to 0 ms) suggests a gradual decay of semantic interference as participants prepare for articulation.

Considering T2T2, it shows different processing from T1T1. At SOA -300 ms, there are no significant effects when distractors have declarative tunes, emphasising T2T2's natural compatibility with interrogative intonation. Unlike T1T1, which requires tonal alternation rule, T2T2 maintains consistent rising contours in both underlying and surface forms, creating a stable prosodic framework that resists interference from competing declarative tunes.

In tune match conditions at this early stage, T2T2 yields semantic inhibition without phonological effects, which suggests that when lexical tones and intonational requirements are naturally aligned, the prosodic system operates efficiently from the start, allowing resources to

be allocated to semantic processing. The semantic inhibition reflects strong activation of the distractor's meaning during early conceptual planning, creating competition that requires resolution.

The continued absence of effects for T2T2 in tune clash conditions at SOA 0 ms reinforces the notion that its rising pattern creates strong resistance to prosodic interference throughout lexical selection. The system appears to prioritise the congruent T2T2-interrogative mapping over any competing prosodic information from declarative distractors.

Significantly, clear phonological facilitation is observed at this timing in tune match conditions, which represents optimal prosodic alignment, in which the distractor's rising interrogative contour perfectly matches both the lexical requirements of T2T2 and the intonational demands of the question. This triple alignment creates processing efficiencies that actively facilitate phonological encoding, providing the clearest evidence that tone-tune compatibility directly influences processing costs during speech planning.

At SOA 300 ms, both tune conditions for T2T2 show significant effects. When distractors have declarative tunes (tune clash conditions), semantic inhibition is found, suggesting that after prosodic planning is largely complete, processing resources become available to handle the distractor's semantic content. The late semantic competition likely reflects monitoring processes that check the selected lexical item against alternatives.

The combination of weak phonological facilitation and semantic inhibition in tune match conditions is found at late production processes. The persistent facilitation suggests continued benefit from pitch alignment during articulation planning, while the semantic inhibition reflects

final monitoring processes where meaning-based competitors remain active. Such dual effect demonstrates how phonological and semantic processes can operate in parallel even during late production stages.

6.5.3 Parallel Processing of Tone and Tune

The present findings have shown a complicated interaction between lexical tone and intonational tune in speech production planning. Our results demonstrate that in tonal languages, these two prosodic systems, lexical tone and intonational tune, are processed in parallel yet interact dynamically across multiple time windows. Unlike in non-tonal languages where intonational patterns can be applied relatively independently (Lahiri, 2009), tonal languages require constant integration and conflict resolution between lexical and sentential prosodic layers.

Our evidence for early tune planning challenges strictly sequential models of speech production (e.g., Levelt, 1989; Levelt et al., 1999), that prosodic frameworks are not merely post-lexical add-ons but are established remarkably early in the planning process, creating constraints that shape subsequent lexical selection. The facilitation effect observed when interrogative tunes in distractors matched the required intonation of targets indicates that tune selection begins during early conceptual planning before lexical access is complete.

6.5.4 Cascade Processing of Language Production

The different patterns observed between T1T1 and T2T2 tone combinations suggest a cascade (overlapped) model of processing, where activation flows continuously between conceptual, lexical, and phonological levels. When lexical tones naturally align with intonational requirements (as in T2T2's rising-rising contour with interrogative tunes), the system operates

efficiently, freeing resources for other processes, evidenced by the shift to semantic effects. When lexical tones conflict with intonational demands (as in T1T1's complex rising-falling pattern), processing resources must be allocated to resolving the prosodic pressure, creating vulnerability to phonological interference, explaining why T2T2 patterns show resistance to prosodic interference at early SOAs and facilitation at 0 ms, while T1T1 patterns demonstrate persistent phonological conflicts across multiple time windows. The cognitive system appears to prioritise resolution of tone-tune conflicts before proceeding with other aspects of speech planning, where multiple phonological requirements must be simultaneously optimised.

Our data also confirm that speakers must simultaneously manage at least three levels of prosodic representation: (1) underlying tonal representations, (2) surface tonal realisations, and (3) intonational tune requirements, which creates potential for multi-level conflicts, particularly evident in T1T1 patterns where tone sandhi creates discrepancies between underlying representations and surface forms.

These representational conflicts are not solved in a single processing stage but persist throughout the planning timeline. The reemergence of phonological inhibition at late SOAs suggests that even during articulation planning, the system remains sensitive to conflicts between competing prosodic representations, which challenges strictly feed-forward models of speech production and supports interactive activation accounts where later processes can influence earlier representations through feedback connections.

An expansion of current production models to account for the temporally overlapping and co-dependent encoding of tone and tune should be proposed. Levelt's framework acknowledges a "prosody generator", while our results suggest this component must be integrated much earlier

in the planning sequence and must manage bidirectional interactions between lexical and sentential prosody.

Thus, we propose that in tonal languages, the production system establishes an abstract prosodic framework at the conceptual message level, setting expectations for pitch trajectory based on communicative intent (statement versus question). As lexical selection unfolds, this framework constrains and is constrained by the tonal requirements of selected words, creating a dynamic system of mutual influence rather than strictly sequential processing. This interactive framework better explains the observed pattern of facilitation and inhibition effects across different SOAs and tone combinations. When tone and tune are compatible, the system can efficiently integrate these prosodic layers, creating processing advantages. When tone and tune conflict, the system must engage in active conflict resolution, consuming resources and creating vulnerability to interference at multiple planning stages. The whole process is exemplified in the chart below with the case from this study:

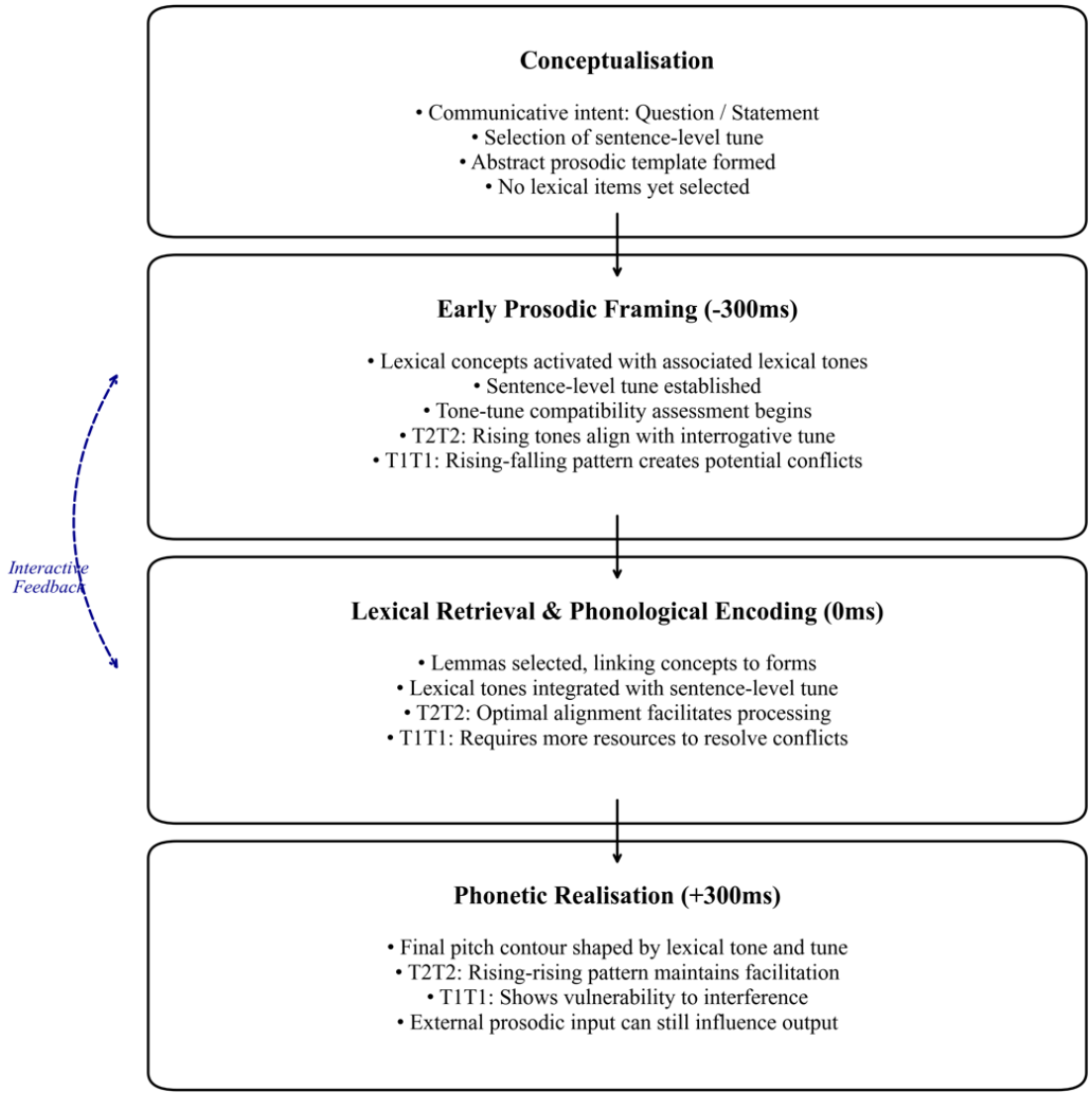


Figure 6.6 Tone-tune processing model in GuanM

6.5.5 Answers to Research Questions

Research Question 1: To what extent is a tune an independent entity in the mental lexicon?

Answer: The results show that tunes possess partial independence in the mental lexicon, while remaining deeply connected with lexical representations. The consistent facilitation effect observed when interrogative tunes in distractors matched the required intonation of targets, even across different SOAs and tone patterns, demonstrates that tunes are accessed as coherent

prosodic units during speech planning, which suggests tunes have representational status distinct from individual lexical items.

However, this independence is constrained by interactions with lexical tones. The differences in processing between T1T1 and T2T2 patterns reveal that tune representation cannot be fully separated from the tonal substrate upon which it is realised. For T2T2 words with rising-rising patterns, the interrogative tune's independence is more pronounced, as evidenced by resistance to declarative tune interference and phonological facilitation with matching tunes. For T1T1 words with rising-falling patterns, the tune's independence appears more compromised, as shown by persistent vulnerability to phonological interference across multiple time windows.

Tunes exist as abstract prosodic templates in the mental lexicon, established early during conceptual planning. However, their realisation and processing are significantly modulated by their compatibility with lexical tonal requirements. Thus, tune independence exists on a gradient, with greater independence observed when tune requirements naturally align with lexical tonal patterns.

Research Question 2: Given that a single word can function as an intonational phrase, how does a tune aligned to a word interact with its phonological (segmental and tonal) and semantic information?

Answer: When a single word functions as an intonational phrase, tune interacts with phonological and semantic information occurs at multiple levels and processing stages:

At the phonological level, tone-tune interactions appear bidirectional. The planned intonational tune creates expectations for particular pitch trajectories, while lexical tones impose constraints

on how these trajectories can be realised. When tune and tone align naturally (as in T2T2 with interrogative tunes), phonological facilitation occurs. When they conflict (as in T1T1 with declarative tune distractors), phonological inhibition results. These effects occurred across different SOAs, indicating continuous integration throughout the planning process.

At the semantic level, tone-tune compatibility influences resource allocation for semantic processing. When tunes and tones aligned well (T2T2 with interrogative tunes), semantic inhibition effects emerged even at early SOAs, suggesting available resources for semantic competition. When resolving tone-tune conflicts required greater processing resources (T1T1 patterns), semantic effects were often reduced or delayed.

Importantly, these interactions differ across the time course of production. Early in planning (SOA -300 ms), tone-tune compatibility appears to establish the basic prosodic framework. During lexical selection (SOA 0 ms), this framework constrains lexical choice and phonological encoding. In late planning (SOA 300 ms), the interaction influences final articulatory adjustments and monitoring processes.

In summary, in single-word intonational phrases, tune does not simply overlay lexical information but interacts with both phonological and semantic properties throughout the planning, which is not static but changes across processing stages reflecting the continuous coordination required to integrate sentence-level prosody with word-level properties.

6.5.6 Limitations

The limitations of the study should be acknowledged, the first of which is the limited number of experimental picture stimuli used in the task. With fewer items, there is an increased risk

that the results were item-specific, rather than reflective of consistent cognitive mechanisms. Second, the tone-related manipulation was asymmetrical across conditions in the stimuli design. To be specific, although the experimental items were designed to involve lexical tone dissimilation rules that produced the surface rising-falling contours, the phonological distractors reflected the underlying falling tone (T1) of the first syllable, since the distractor's tone was based on its underlying representation, whereas the speaker's target involved the surface tone realisation. This mismatch may have driven phonological inhibition effects that were less attributable to tone-tune interaction and more likely a result of morphophonological incongruence locally. As a result, this study cannot fully separate surface tonal conflict from underlying tonal activation during planning. Therefore, future studies should ensure that the surface tone of the target is reflected in the phonological form of the distractor. In this case, it would be through the inclusion of T2. Third, although semantic distractors were evaluated for semantic closeness, their tone combinations were not controlled. Therefore, it remains unclear whether some observed semantic inhibition effects might have been modulated by covert phonological or prosodic overlap, especially in T2T2 conditions, where tone-tune alignment was stronger. Future work could control for tone contour even in semantically driven trials to fully disentangle semantic and prosodic competition.

6.6 Conclusion

This study is the first investigation of intonational tune planning in a tonal language using a PWI paradigm. By applying this well-established psycholinguistic method, we have demonstrated how speakers coordinate lexical tone with sentence-level intonation during real-time speech production, from initial planning through to articulation.

The observed effects across SOAs show a dynamic interaction between lexical tone and intonational tune that evolves throughout the production timeline. In this experiment, participants produced interrogative utterances while being exposed to auditory distractors that either matched or mismatched the intended intonational tune. These distractors varied in semantic or phonological relatedness to picture stimuli and were presented at three different SOAs: -300 ms (before picture onset), 0 ms (simultaneous with onset), and 300 ms (after onset). This design allowed us to examine precisely how auditory prosodic input interacts with internal planning at different processing stages.

Our finding of faster reaction times when distractors carried an interrogative tune confirms that intonational tunes function as abstract prosodic entities that are accessed early during conceptual planning, rather than being mere post-lexical add-ons. This early tune selection subsequently constrains phonological encoding, creating a framework within which lexical tones must be integrated. The bidirectional nature of this interaction is evident in how the shape of lexical tones (T1T1: rising-falling; T2T2: rising-rising) influences not only phonological planning but also modulates the processing resources allocated to resolving potential conflicts between lexical and sentential prosody. However, the tone-tune integration is highly sensitive to the phonological structure of lexical tones themselves. T1T1 and T2T2 patterns show different effects of inhibition and facilitation since lexical tone does not simply coexist with intonation but must be actively integrated with it through a dynamic process of conflict resolution or reinforcement, depending on their inherent compatibility with tunes. Thus, T2T2 tones, with their rising-rising contour, aligned naturally with interrogative tunes and facilitated production, especially when distractors also carried rising pitch. In contrast, T1T1 sequences with their complex rising-falling surface pattern showed increased interference, particularly when faced with competing prosodic information. This asymmetry suggests that the production

system must manage multiple levels of prosodic representation simultaneously including underlying tones, surface tones, and intonational tunes at any level capable of disrupting the planning process.

In summary, this study demonstrates that the interaction between lexical tone and intonational tune is not peripheral or post-lexical but deeply embedded within the speech planning architecture. The early, abstract representation of tunes and their continuous interaction with lexical tones throughout the production timeline challenges strictly sequential models of speech production and suggests a more interactive, resource-sharing framework. By extending experimental research on tune planning to a tonal language, this research study hopes to contribute to the understanding of how prosodic systems, both lexical and phrasal, are synchronised in real time. Future studies might profitably expand on these findings by investigating longer utterances and cross-dialectal comparisons to determine how widespread these planning patterns are across diverse tonal and intonational systems.

Chapter 7 Final Discussions

This chapter begins with a review of the key findings from four studies presented in the thesis, followed by a broader discussion of tone-tune interaction, as well as the applications and limitations of the AM framework in relation to the studies. The chapter concludes by outlining directions for future research and reflecting on the overall contributions of the thesis.

7.1 Summaries of Experiments

Chapters 3 and 4 explored the production of syntactically unmarked yes-no questions in GuanM, both in isolation and within longer utterances containing prosodic focus. The target items were disyllabic words, including one minimal pair involving a tonal dissimilation rule. The aim was to explore how intonation is realised in a tonal language, and how it interacts with lexical tone and sentence-level features such as focus.

The findings reveal that:

- Statements are characterised by a global low pitch register, which gives rise to a gradual F0 declination across the utterance. This declination is not driven by tonal structure, but is interpreted as the result of the intonational register setting.
- In contrast, yes-no questions exhibit a global high pitch register and a final high boundary tone, aligned to the right edge of the intonational phrase. This boundary tone has different phonetic realisations on the final syllable: it enhances the rise of rising tones, moderates the fall of falling tones, and converts level tones into rising contours.
- In longer utterances, the combination of high register and boundary tone produces a smoother and more level F0 contour compared to the downward-sloping contour found in statements. Importantly, lexical tone contours, including those affected by tonal dissimilation, remain intact, although their pitch range may be varied. These findings

suggest that in GuanM, contour shape is more phonologically stable than register height. While register may be modulated by sentence type, the tonal identity of lexical items is preserved.

- Focus is primarily marked by shorter duration (reflecting faster speech rate) and increased intensity, rather than by F0-based cues. In statements, focus can also be marked by an H* pitch accent. However, in questions, the interaction between lexical tone and the question tune sometimes weakens the realisation of H*, making it a less reliable cue.

Chapter 5 investigated the interaction between lexical tone and intonational tune through a perception experiment using the AX discrimination paradigm. The results show that perceptual difficulty increases when the pitch trajectories of lexical tone and intonational tune are congruent, particularly when both exhibit rising contours. In such cases, listeners struggle to distinguish whether the pitch movement arises from lexical tone or the interrogative intonation, leading to increased ambiguity. In contrast, no comparable effect is found when the pitch trajectories are incongruent, such as when falling tones co-occur with rising question tunes or when falling tones align with falling declarative tunes. These findings highlight an asymmetry in tone-tune interaction, supporting the conclusion that contour shape, rather than pitch register, is the primary perceptual cue distinguishing question and statement tunes in GuanM.

Chapter 6 examined intonational tune planning (particularly the interrogative tune) in GuanM using a picture-word interference paradigm. In this experiment, participants were asked to name pictures using either interrogative or declarative tunes, while auditory distractors were presented with either interrogative or declarative intonation at three SOAs. The results support a facilitation effect when producing interrogative tunes in the presence of interrogative-tune

distractors, suggesting that tune planning occurs at an early stage of speech production. Further analysis shows an asymmetrical result of phonological and semantic effects between different tone combinations. T2T2 words with rising-rising contours showed natural alignment with interrogative tunes, creating processing advantages and resistance to interference from competing declarative tunes. In contrast, T1T1 words with rising-falling patterns demonstrated greater vulnerability to prosodic competition. These findings challenge modular models of speech production by revealing that intonational tune is not merely a post-lexical add-on but functions as an abstract prosodic entity that is accessed early in conceptual planning and interacts dynamically with lexical tone throughout the production process. As a result, speakers of tonal languages must simultaneously manage multiple layers of prosodic representation, including underlying tones, surface tones, and intonational tunes.

7.2 General Discussion

7.2.1 Tone-Tune Interaction

The theoretical question of whether intonational tunes are abstract, phonological entities that are planned and integrated into speech from the earliest stages, particularly within tonal languages (e.g., GuanM) in which pitch is already lexically contrastive, was investigated in this thesis. Evidence from the four studies reported in this thesis provides strong support for the view that intonational tunes are structurally represented and that lexical tone and intonation interact across all levels of spoken language, ranging from message-level planning to articulation and perception.

In tonal languages, pitch has two grammatical functions; it encodes both word-level lexical contrasts and utterance-level prosodic meaning. This duality challenges traditional models of intonation that treat tunes as post-lexical phonetic overlays that are added after the core

linguistic content is specified. By contrast, the results of the study of GuanM demonstrate that intonational tunes must be represented at an abstract level and interact with lexical tone throughout the speech planning and realisation processes.

In the earliest planning stage, the PWI study (Chapter 6) revealed that intonational tunes were accessed during conceptual message formulation, well before phonological encoding was complete. The facilitation effect that occurred when the distractors had matching interrogative tunes suggests that the speakers selected an abstract tune template based on communicative intent (question versus statement) that may have shaped subsequent lexical selection. This early planning of tunes challenges strictly sequential models of speech production and shows that intonational categories have representational status independent of specific lexical content. As the planning progressed to phonological encoding, the planning progressed differently when different lexical tones were involved. T2T2 words with rising-rising contours showed a natural alignment with interrogative tunes, thereby creating processing advantages, while T1T1 words with complex rising-falling patterns revealed vulnerability to prosodic competition.

During articulation, the production studies (Chapters 3-4) showed how this abstract tune was realised physically. Questions were consistently realised through global pitch raising across the entire utterance, coupled with a final high boundary tone, while statements had a global low register. The raised pitch register of questions was not limited to short phrases; instead, it was observed across both disyllabic targets and longer utterances, indicating that the interrogative tune was not a localised rising contour, but was encoded through the phonological intonational structure. Furthermore, the intonational features were implemented without compromising the lexical tone identity; even tones that were affected by dissimilation rules maintained their contour shapes, albeit with adjustments to pitch range.

The perception data further confirmed that the listeners treated intonational tunes as structured, meaningful units. When the lexical tone and tune shared pitch features, particularly in rising-rising combinations, the listeners' discrimination accuracy decreased significantly, suggesting perceptual interference due to acoustic overlap. The low level of accuracy would not have appeared if intonation were only processed after the lexical tone, or if it lacked categorical status. This shows that tone and tune were parsed together, requiring listeners to disambiguate overlapping pitch cues in real time.

From planning through to production and perception, the findings consistently showed that intonational tunes in tonal languages functioned as abstract prosodic templates that were activated early in speech formulation and interacted progressively with lexical tones throughout the cognitive process.

7.2.2 Applications and Limitations of AM Theory

AM Theory has demonstrated some success in modelling pitch patterns in GuanM, particularly in the case of yes-no questions, through a binary representation of high and low tones. It offers a relatively accessible method for visualising surface-level pitch contours and has proven to be useful in describing both lexical tones and intonational melodies in a simplified manner.

However, the theory's ability to model the phonological components of GuanM is limited, particularly with regard to the interaction between lexical tone and intonation. In tonal languages such as GuanM, lexical tone is not simply an underlying layer upon which intonation is superimposed, but plays an active and equally significant role in shaping the prosodic

structure of an utterance. There is a complex interaction between lexical tones and sentential intonation throughout the stages of speech production, perception and planning.

Therefore, there is a strong argument for representing lexical tones on a separate tonal tier within the AM framework. Instead of viewing tone as a foundation and intonation as a higher-level supervisor, a more accurate characterisation would be to regard them as two interconnected layers, each of which is essential, with each contributing to the overall phonetic and phonological outcome. Tone does not passively provide a base for intonation, but actively shapes the possible pitch movements and the perception of intonational patterns.

In this thesis, I argued that the interaction between tone and tune is not unidirectional - that is, with intonation affecting tones - but is bidirectional, with lexical tones also influencing intonation. For example, in the production study of short utterances, the tone of the final syllable determined the phonetic realisation of the high boundary tone introduced by intonation: If the final lexical tone was rising, the boundary tone also rose; if it was falling, the boundary tone reflected that fall. In turn, the boundary tone can exaggerate the rising of a rising tone or reduce the falling nature of a falling tone. In longer utterances, the global register raising associated with the interrogative intonation reduced the contour shape of each lexical tone, resulting in a more level overall pitch contour with fewer fluctuations. This reduction in pitch variability contrasts with the findings for Neapolitan Italian (Cangemi & Grice, 2016), in which interrogatives exhibited significantly greater F0 variability compared to declaratives regardless of the sentence focus or duration. Thus, the advantage of incorporating both tonal and intonational tiers is mutual predictability: Intonational patterns become more predictable based on the lexical tones, while the surface realisation of lexical tones becomes more consistent when influenced by intonation.

AM Theory also encounters difficulty in accounting for the global tendencies that are observed in intonational patterns because it relies heavily on identifying specific turning points in the pitch contour. The timing and scaling of H and L tone sequences are presumed to be sufficient to signal tonal contrasts. However, although turning points are often ambiguous or absent in natural speech, listeners remain capable of making clear judgements about pitch accents, which strongly suggests that other factors, such as the overall pitch contour shape and distribution, play crucial roles in tonal perception. By focusing narrowly on local targets while disregarding broader contour characteristics, AM Theory fails to capture the full range of perceptual cues on which speakers and listeners rely in authentic communication (Barnes et al., 2012). In addition, the empirical data obtained from GuanM show that the observed global tendency, including the global register raising in question tunes and the global low register with gradual declination in statement tunes, cannot be adequately explained by the local target-based representations of AM Theory.

To address this, an extension of the AM framework should incorporate a mechanism for global pitch register tracking in GuanM and other Chinese dialects, which would involve modelling not just the sequence of H and L tones, but also a higher-level parameter that defines the overall pitch range or slope of the utterance. Such a parameter could represent a rising register trend in interrogatives or a declination pattern in declaratives, for example. It could take the form of a register tier that is applied across the intonational phrase, or a scaling factor that modulates the height of local tones in accordance with the global contour. This approach would retain the strengths of the AM model in identifying local pitch targets while enhancing its ability to represent the long-range pitch patterns that listeners demonstrably perceive. Crucially, it would

also enable the theory to better account for intonational phenomena in tonal languages, in which such global contours frequently interact with the F0 realisation of lexical tones.

7.3 Future Studies

The focus in this thesis was on disyllabic sequences, which served as experimental units across all the studies. In Chapter 3, these disyllabic sequences functioned as complete intonational phrases; in Chapter 4, they were embedded in focus positions within longer utterances and, in Chapter 5, they formed the perceptual stimuli for discrimination tasks. In Chapter 6, both distractors and target productions were disyllabic sequences that functioned as intonational phrases. While this method provided a controlled environment for examining tone-tune interactions, natural speech typically involves more complex utterances. Future research should examine longer utterances to determine whether tone-tune interaction in GuanM is organised hierarchically across prosodic boundaries. Such work should analyse multiple acoustic dimensions (duration, intensity, and F0 measures) and draw on larger datasets, ideally including comparative data from other Chinese languages.

The comparison between syntactically marked and unmarked yes-no questions presents another important aspect for investigation. The current findings demonstrated a global raised pitch register in syntactically unmarked yes-no questions in both disyllabic utterances and in longer utterances, suggesting that speakers plan this intonational pattern from the beginning of the utterance. Shen's (1990) work on yes-no questions in SC indicated that syntactically marked yes-no questions occupied an intermediate position in terms of register height, with a mean F0 that was higher than that of statements but lower than that of intonational yes-no questions. This confirms that intonational tune exists as an abstract phonological entity in which global pitch register is embedded, even for the sentence-final particles in yes-no

questions. Such observations raise intriguing questions about the systematic relationship between the syntactic and prosodic marking of interrogatives in GuanM, particularly regarding whether the global high pitch register, representing the abstract entity of interrogative tune, will also be found in these contexts.

In addition, the speech rate in the current findings is a particularly promising area for future investigations. The observation that focus was primarily marked by a shorter duration rather than by F0 modulation suggests that temporal features play a more significant role in GuanM prosody than was previously recognised. This makes sense given that F0 is heavily utilised for lexical tone distinctions, potentially limiting its availability for marking information-structural categories like focus. Future research should examine whether this pattern is maintained in longer focus phrases with more syllables. Would speakers compress the entire focused phrase to maintain a faster rate, or would they shift to alternative prosodic cues? Moreover, how might the speech rate interact with tonal and intonational features in such contexts?

Finally, apart from the recommendations outlined in Chapter 6, future research on tune planning in speech production could be suggested. In the present study, the PWI paradigm was used to investigate the production of question and statement tunes in two-syllable words functioning as intonational phrases, using measures of reaction times and accuracy. While this approach provided evidence for the early planning of tune and confirmed the interaction between lexical access and prosodic realisation, recent work by Ots (2024) has offered a broader perspective by examining the planning of longer, sentence-level utterances. Ots' investigation of Estonian intonation revealed that sentence-initial pitch peaks were modulated by utterance length, suggesting that speakers engaged in the advance planning of intonational contours, which was not linked to the incrementality of conceptual planning, as measured by

eye-tracking, but was linked to the structural features of the utterance. She suggested that prosodic features, such as pitch scaling, might not be solely linked to lexical retrieval or to local phonological processes, but may instead be informed by higher-level syntactic or discourse structures.

Drawing on Ots' (2024) study, future psycholinguistic research on tune planning could adopt a dual-task paradigm, in which participants produce intonational tunes while simultaneously managing cognitive load, for example, by memorising a list of words. This would allow researchers to examine how attentional resources affect the planning and realisation of different tune types, such as questions and statements. In addition, the simple picture-naming task that was used in the current study could be replaced with a visual world speech production paradigm involving scene descriptions, which would allow for the study of tune planning in more naturalistic and discourse-rich contexts, in which participants are required to generate full utterances. Scene descriptions might be useful to explore sentence-level or phrasal contexts to investigate how tune planning operates beyond isolated words, by eliciting both questions and statements; for example, through dialogue prompts or information-seeking scenarios. Moreover, instead of focusing solely on reaction times and accuracy, future work could incorporate acoustic analyses of F0 trajectories across the entire utterance. This would enable a more detailed examination of how initial pitch scaling contributes to overall tune realisation, and could determine whether intonation is planned holistically or incrementally across different tune types. Finally, eye-tracking methodology could be implemented to capture real-time planning processes: the timing of tune selection could be determined by analysing measures such as such as gaze durations and fixation patterns prior to speech onset. Such data could help to clarify whether question and statement tunes are planned differently in terms of scope and temporal structure.

7.4 Conclusion

In summary, this thesis has investigated the intonational system of GuanM. It has found that syntactically unmarked yes-no questions in GuanM exhibit a raised global register and a high boundary tone at the edge of the intonational phrase, and has argued that intonational tunes are abstract entities that are structurally represented, perceptually salient, and fully embedded in the architecture of language use. These tunes interact with lexical tones across multiple levels, from conceptualisation to articulation to perception. These findings call for a revision of current models of prosody, particularly those based on data from non-tonal languages since in tonal languages, pitch serves as a multifunctional resource.

The findings in this thesis make several significant theoretical contributions to our understanding of the interaction between lexical tone and intonation. Of note, they challenge the traditional view that intonational meaning in tonal languages is primarily realised through localised prosodic adjustments. Instead, I have demonstrated that GuanM employs a global strategy wherein speakers initiate a raised pitch register from the onset of interrogative utterances, reflecting the advanced planning of the entire intonational contour. This planning occurs simultaneously with the preservation of lexical tone contrasts.

The methodology that was used in this thesis combined acoustic analyses and perceptual research to provide a comprehensive and formal framework for studying the intonation systems of less-studied languages. By examining production and perception, I was able to gather evidence for the cognitive representations of intonation categories in GuanM, suggesting that speakers and listeners have consistent expectations about how the prosodic encoding of interrogative meanings should be accomplished. In addition, I explored GuanM speakers' psycholinguistic processing of the intonation planning of a multilayered prosodic system. The

GuanM speakers' seemingly effortless management of the complex interaction between lexical tone and intonation raises interesting questions about the structure of speech planning and the nature of phonological representations.

In conclusion, I hope that this thesis will serve as a useful template for future research on less-documented languages, particularly those that have historically relied on impressionistic descriptions rather than empirical evidence, by using diverse research methods that were covered in the thesis, including production with acoustic measurements, perception and psycholinguistic experiments. I also hope this work will encourage similarly comprehensive methodologies in the study of other tonal languages, which will enhance our understanding of tone-tune interaction.

References

Alderete, J., Chan, Q., & Yeung, H. H. (2019). Tone slips in Cantonese: Evidence for early phonological encoding. *Cognition*, *191*, 103952. https://doi.org/10.1016/j.cognition.2019.04.021

Anderson, S. R. (1976). Nasal consonants and the internal structure of segments. *Language*, *52*(2), 326–344. https://doi.org/10.2307/412563

Armstrong, M. E. (2017). Accounting for intonational form and function in Puerto Rican Spanish polar questions. *Probus*, *29*(1), 1–40. https://doi.org/10.1515/probus-2014-0016

Arvaniti, A. (2011). The representation of intonation. In M. Oostendorp, E. Hume, C. J. Ewen, & K. Rice (Eds), *The Blackwell companion to phonology* (Vol. 5, pp. 757–780). Wiley-Blackwell.

Arvaniti, A. (2022). The Autosegmental-Metrical model of intonational phonology. In J. Barnes & S. Shattuck-Hufnagel (Eds), *Prosodic theory and practice* (pp. 25–63). The MIT Press. https://doi.org/10.7551/mitpress/10413.003.0004

Arvaniti, A., & Baltazani, M. (2005). Intonational analysis and prosodic annotation of Greek spoken corpora. In S.-A. Jun (Ed.), *Prosodic typology* (pp. 84–117). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0004

Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics*, *26*(1), 3–25. https://doi.org/10.1006/jpho.1997.0063

Bao, Z. (1999). *The structure of tone*. Oxford University Press.

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2). https://doi.org/10.1515/lp-2012-0017

Baxter, W. H. (1992). *A Handbook of Old Chinese Phonology*. De Gruyter Mouton. https://doi.org/10.1515/9783110857085

Baxter, W. H., & Sagart, L. (2014). *Old Chinese: A new reconstruction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199945375.001.0001

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309. https://doi.org/10.1017/S095267570000066X

Beijing Daxue Zhongguo Yuyan Wenxue Xi Yuyanxue Jiaoyanshi (Ed.). (2003). *Hanyu fangyin zihui [A collection of Chinese dialect pronunciations]* (2nd edn). Yuwen Chubanshe.

Bibyk, S. A. (2016). A rise by any other name: An investigation of the production and comprehension of rising (and falling) intonation in questions [Ph.D., University of Rochester]. In *ProQuest Dissertations and Theses* (1858816375). ProQuest Dissertations & Theses Global; ProQuest One Literature; Social Science Premium Collection. https://www.proquest.com/dissertations-theses/rise-any-other-name-investigation-production/docview/1858816375/se-2?accountid=13042

Bolinger, D. L. (1951). Intonation: Levels versus configurations. *Word*, *7*(3), 199–210. https://doi.org/10.1080/00437956.1951.11659405

Borràs-Comes, J., Vanrell, M. D. M., & Prieto, P. (2014). The role of pitch range in establishing intonational contrasts. *Journal of the International Phonetic Association*, *44*(1), 1–20. https://doi.org/10.1017/S0025100313000303

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Gleerup.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. *Laboratory Phonology*, *7*(1), 9. https://doi.org/10.5334/labphon.28

Chang, M.-C. L. (1992). A prosodic account of tone, stress, and tone sandhi in Chinese languages [Ph.D., University of Hawai'i at Manoa]. In *ProQuest Dissertations and Theses* (303974798). ProQuest Dissertations & Theses Global; ProQuest One Literature. https://www.proquest.com/dissertations-theses/prosodic-account-tone-stress-sandhi-chinese/docview/303974798/se-2?accountid=13042

Chang, N.-C. T. (1958). Tones and intonation in the Chengtu dialect (Szechuan, China). *Phonetica*, *2*(1–2), 59–85. https://doi.org/10.1159/000257848

Chao, Y. R. (1930). Ə sistim əv 'toun-letəz'. *Le Maître Phonétique*, *8 (45)*(30), 24–27.

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology, Academia Sinica*, *4*(2), 121–134.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chappell, H. (2001). Synchrony and diachrony of Sinitic languages: A brief history of Chinese dialects. In H. Chappell (Ed.), *Sinitic grammar: Synchronic and diachronic perspectives* (pp. 3–27). Oxford University Press.

Chen, J.-Y. (1999). The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue. *Applied Psycholinguistics*, *20*(2), 289–301. https://doi.org/10.1017/S0142716499002064

Chen, Y. (2010). Post-focus F0 compression—Now you see it, now you don't. *Journal of Phonetics*, *38*(4), 517–525. https://doi.org/10.1016/j.wocn.2010.06.004

Cheng, H.-S., Niziolek, C. A., Buchwald, A., & McAllister, T. (2021). Examining the relationship between speech perception, production distinctness, and production variability. *Frontiers in Human Neuroscience*, *15*, 660948. https://doi.org/10.3389/fnhum.2021.660948

Cheng, R. L. (1966). Mandarin phonological structure. *Journal of Linguistics*, *2*(2), 135–158.

Chłopicki, W. (2019). Declarative questions in Polish student conversations. *Journal of Pragmatics*, *153*, 69–79. https://doi.org/10.1016/j.pragma.2019.03.003

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Connell, B. (2001, May 18). Downdrift, downstep, and declination. *Proceedings of Typology of African Prosodic Systems Workshop*. Typology of African Prosodic Systems Workshop, Bielefeld University, Germany.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142–2156. https://doi.org/10.1121/1.392372

Damian, M. F., & Bowers, J. S. (2003). Locus of semantic interference in picture-word interference tasks. *Psychonomic Bulletin & Review*, *10*(1), 111–117. https://doi.org/10.3758/BF03196474

Damian, M. F., & Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 345–361. https://doi.org/10.1037/0278-7393.25.2.345

De Rosario-Martinez, H. (2012). *phia: Post-hoc interaction analysis* (p. 0.3-1) [Data set]. https://doi.org/10.32614/CRAN.package.phia

Dell, F. (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst, & J.-R. Vergnaud, *Forme sonore du langage: Structure des représentations en phonologie* (pp. 65–122). Hermann.

Downing, L. J., & Rialland, A. (2016). Introduction. In L. J. Downing & A. Rialland (Eds), *Intonation in African tone languages* (pp. 1–16). De Gruyter. https://doi.org/10.1515/9783110503524-001

Duan, W., & Jia, Y. (2014). The typology of focus realization of Northern Mandarin. *The 9th International Symposium on Chinese Spoken Language Processing*, 492–496. https://doi.org/10.1109/ISCSLP.2014.6936589

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd edn). Oxford University Press.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *The Journal of the Acoustical Society of America*, *80*(2), 402–415. https://doi.org/10.1121/1.394091

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, *36*(4), 680–703. https://doi.org/10.1016/j.wocn.2008.05.001

Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 39–55). Springer New York. https://doi.org/10.1007/978-1-4613-8202-7_3

Fujisaki, H. (2004). Information, prosody, and modeling—With emphasis on tonal features of speech -. *Speech Prosody 2004*, 1–10. https://doi.org/10.21437/SpeechProsody.2004-1

Gårding, E. (1983). A generative model of intonation. In A. Cutler & D. R. Ladd (Eds), *Prosody: Models and measurements* (Vol. 14, pp. 11–25). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69103-4_2

Gårding, E. (1987). Speech act and tonal pattern in Standard Chinese: Constancy and variation. *Phonetica*, *44*(1), 13–29. https://doi.org/10.1159/000261776

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–376. https://doi.org/10.3758/BF03194885

Glaser, W. R., & Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 640–654. https://doi.org/10.1037/0096-1523.10.5.640

Goldsmith, J. A. (1976). *Autosegmental phonology*. Massachusetts Institute of Technology.

Gong, Q. (1995). Guanzhong fangyan de biandiao he bianyin [Tone sandhi and sound changes in the Guanzhong dialect]. *Yuwen Yanjiu*, *4*, 55–58.

Grabe, E., Kochanski, G., & Coleman, J. (2003). Quantitative modelling of intonational variation. *Proceedings of Speech Analysis and Recognition in Technology, Linguistics and Medicine*.

Gussenhoven, C. (Ed.). (2004). Paralinguistics: Three biological codes. In *The Phonology of tone and intonation* (pp. 71–96). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511616983.006

Gussenhoven, C., & Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, vols 2, 91-94–0. https://doi.org/10.21437/ICSLP.2000-216

Gussenhoven, C., & Rietveld, T. (2000). The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech*, *43*(2), 183–203. https://doi.org/10.1177/00238309000430020301

Gussenhoven, C., & Van de Ven, M. (2020). Categorical perception of lexical tone contrasts and gradient perception of the statement–question intonation contrast in Zhumadian Mandarin. *Language and Cognition*, *12*(4), 614–648. https://doi.org/10.1017/langcog.2020.14

Haan, J. (2002). *Speaking of questions: An exploration of Dutch question intonation*. Netherlands Graduate School of Linguistics.

Hart, J. T., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody* (1st edn). Cambridge University Press. https://doi.org/10.1017/CBO9780511627743

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language & Linguistic Theory*, *9*(1), 47–96. https://doi.org/10.1007/BF00133326

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, *31*(1), 39–62. https://doi.org/10.1016/S0095-4470(02)00071-2

Hennoste, T., Rääbis, A., & Rumm, A. (2019). Estonian declarative questions: Their usage and comparison with vä- and jah-questions. *Journal of Pragmatics*, *153*, 46–68. https://doi.org/10.1016/j.pragma.2019.04.010

Henriksen, N. C. (2010). Question intonation in Manchego Peninsular Spanish [Ph.D., Indiana University]. In *ProQuest Dissertations and Theses* (725641648). ProQuest Dissertations & Theses Global; ProQuest One Literature; Social Science Premium Collection. https://www.proquest.com/dissertations-theses/question-intonation-manchego-peninsular-spanish/docview/725641648/se-2?accountid=13042

Hirst, D., & Di Cristo, A. (1998). *Intonation systems: A survey of twenty languages*. Cambridge University Press.

Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (Ed.), *Prosody: Theory and experiment* (Vol. 14, pp. 51–87). Springer Netherlands. https://doi.org/10.1007/978-94-015-9413-4_4

Ho, A. T. (1977). Intonation variation in a Mandarin sentence for three expressions: Interrogative, exclamatory, and declarative. *Phonetica*, *34*(6), 446–457. https://doi.org/10.1159/000259916

Hockett, C. F. (1955). A *manual of phonology*. Waverly Press.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press.

Huang, J., & Zhang, G. (2019). A study on prosodic distribution of yes/no questions with focus in Mandarin. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds), *Natural language processing and Chinese computing* (Vol. 11838, pp. 565–580). Springer International Publishing. https://doi.org/10.1007/978-3-030-32233-5_44

Hui, Z., Hong, J., Lu, Y., Ru, J., & Wang, L. (2020). Xi'an putonghua cihui he yufa tedian diaocha yanjiu [A study on the vocabulary and grammatical features of Xi'an Putonghua]. *Pinwei Jingdian*, *11*, 17–20.

Humphreys, G. W., Besner, D., & Quinlan, P. T. (1988). Event perception and the word repetition effect. *Journal of Experimental Psychology: General*, *117*(1), 51–67. https://doi.org/10.1037/0096-3445.117.1.51

Hyman, L. M., & Monaka, K. C. (2011). Tonal and non-tonal intonation in Shekgalagari. In S. Frota, G. Elordieta, & P. Prieto (Eds), *Prosodic categories: Production, perception and comprehension* (pp. 267–289). Springer Netherlands. https://doi.org/10.1007/978-94-007-0137-3_12

Jiang, H. (2019). Intonational pitch features of interrogatives and declaratives in Chengdu dialect. *Forum for Linguistic Studies*, *1*(1), 77. https://doi.org/10.18063/fls.v1i1.1082

Jiang, P., & Chen, A. (2011). Representation of Mandarin intonations: Boundary tone revisited. *The 23rd North American Conference on Chinese Linguistics (NACCL-23)*, *1*, 97–109.

Jones, D. (1976). *An outline of English phonetics* (9th ed.). Cambridge University Press.

Keating, P., & Shattuck-Hufnagel, S. (2002). A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics*, *101*, 112–156.

Kurpaska, M. (2010). *Chinese language(s): A look through the prism of the Great dictionary of modern Chinese dialects*. Mouton de Gruyter.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonology Yearbook*, *1*, 53–74. https://doi.org/10.1017/S0952675700000294

Ladd, D. R. (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology Yearbook*, *3*, 311–340. https://doi.org/10.1017/S0952675700000671

Ladd, D. R. (1996). *Intonational phonology*. Cambridge University Press.

Ladd, D. R. (2008). *Intonational phonology* (2nd ed). Cambridge University Press.

Lahiri, A. (2009). *Intonational and tonal lexicon* [Conference presentation]. International Workshop on Spoken Language Prosody (IWSLPR-09), Kolkata, India.

Lan, B. (2011). *Xi'an fangyan yufa diaocha yanjiu [A grammatical investigation and study of the Xi'an dialect]*. Zhonghua Shuju.

Leben, W. R. (1973). *Suprasegmental phonology*. Massachusetts Institute of Technology.

Lenth, R. V. (2024). *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT press.

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, *42*(1–3), 1–22. https://doi.org/10.1016/0010-0277(92)90038-J

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01). https://doi.org/10.1017/S0140525X99001776

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, P. (2001). The correspondence pattern between Xi'an dialect and Putonghua in pronunciation. *Journal of Xi'an Educational College*, *2*, 57–61.

Liang, J., & Heuven, V. J. (2007). Chinese tone and intonation perceived by L1 and L2 listeners. In C. Gussenhoven & T. Riad (Eds), *Tones and Tunes: Experimental Studies in Word and Sentence Prosody* (Vol. 2, pp. 27–62). Mouton de Gruyter. https://doi.org/10.1515/9783110207576.1.27

Liberman, M. Y. (1975). *The intonational system of English* [Ph.D.]. Massachusetts Institute of Technology.

Liberman, M. Y., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249–336.

Lin, E., Daniell, P., & Al Busaidi, S. (2013). A cross-language comparison of pitch patterns in declarative questions and statements. *Speech, Language and Hearing*, *16*(3), 119–126. https://doi.org/10.1179/2050571X13Z.00000000014

Lin, M. (2004). *On production and perception of boundary tone in Chinese intonation*. 125–130. https://www.isca-archive.org/tal_2004/lin04_tal.html#

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62*(2–4), 70–87. https://doi.org/10.1159/000090090

Liu, F., & Xu, Y. (2007). The neutral tone in question intonation in Mandarin. *Interspeech 2007*, 630–633. https://doi.org/10.21437/Interspeech.2007-276

Liu, M., Chen, Y., & Schiller, N. O. (2020). Tonal mapping of Xi'an Mandarin and Standard Chinese. *The Journal of the Acoustical Society of America*, *147*(4), 2803–2816. https://doi.org/10.1121/10.0000993

Liu, M., Chen, Y., & Schiller, N. O. (2022). Context matters for tone and intonation processing in Mandarin. *Language and Speech*, *65*(1), 52–72. https://doi.org/10.1177/0023830920986174

Liu, Z., Van De Velde, H., & Chen, A. (2018). Intonational realization of declarative questions in Bai. *Speech Prosody 2018*, 124–128. https://doi.org/10.21437/SpeechProsody.2018-25

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2006a). Effect of intonation on Cantonese lexical tones. *The Journal of the Acoustical Society of America*, *120*(6), 3978–3987. https://doi.org/10.1121/1.2363927

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2006b). Perception of intonation in Cantonese. *Tonal Aspects of Languages*.

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2011). The perception of intonation questions and statements in Cantonese. *The Journal of the Acoustical Society of America*, *129*(2), 1012–1023. https://doi.org/10.1121/1.3531840

Ma, M. (2005). Acoustic study of the tones of Xi'an dialect. *Journal of Yan'an University (Social Science Edition)*, *27*(4), 110–112.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory*. Psychology Press. https://doi.org/10.4324/9781410611147

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 555–563. https://doi.org/10.1037/0278-7393.28.3.555

Ni, J., & Kawai, H. (2004). Pitch targets anchor Chinese tone and intonation patterns. *Speech Prosody 2004*, 95–98. https://doi.org/10.21437/SpeechProsody.2004-22

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments

in English. *Cognition*, *115*(2), 282–302. https://doi.org/10.1016/j.cognition.2010.01.001

Ots, N. (2024). Planning sentences and sentence intonation in Estonian. *Laboratory Phonology*. https://doi.org/10.16995/labphon.8722

Patin, C. (2016). Tone and intonation in Shingazidja. In L. J. Downing & A. Rialland (Eds), *Intonation in African tone languages* (pp. 285–320). De Gruyter. https://doi.org/10.1515/9783110503524-009

Paul Boersma, & Weenink, D. (2023). *Praat: Doing phonetics by computer [Computer program]. Version 6.3.16*. http://www.praat.org/

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

Peng, S., Chan, M. K. M., Tseng, C., Huang, T., Lee, O. J., & Beckman, M. E. (2005). Towards a Pan-Mandarin system for prosodic transcription. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (1st edn, pp. 230–270). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0009

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* [Ph.D.]. Massachusetts Institute of Technology.

Pierrehumbert, J. B. (2022). Commentary on Chapter 11: Comparing the PENTA model to autosegmental-metrical phonology. In J. Barnes & S. Shattuck-Hufnagel (Eds), *Prosodic theory and practice* (pp. 408–424). The MIT Press. https://doi.org/10.7551/mitpress/10413.003.0021

Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese tone structure*. MIT Press.

Pike, K. L. (1945). *The intonation of American English*. University of Michigan Press.

Qi, W. (2011). Discussion of “pf” and “pfh” in Xi’an dialect. *Journal of Xianyang Normal University*, *26*(1), 73–74.

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsey, S. R. (1987). *The languages of China*. Princeton University Press.

Ren, H., & Yang, N. (2012). Xi’an diqu Putonghua yu Shaanxi hua de shehui gongneng diaocha yu jiedu [A sociolinguistic investigation and interpretation of the social functions of Mandarin and Shaanxi dialect in the Xi’an region]. *Journal of Northwest University (Philosophy and Social Sciences Edition)*, *42*(3), 180–182.

Rialland, A. (2009). Question prosody: An African perspective. In T. Riad & C. Gussenhoven (Eds), *Volume 1 Typological Studies in Word and Sentence Prosody* (pp. 35–62). De Gruyter Mouton. https://doi.org/10.1515/9783110207569.35

Rialland, A., & Aborobongui, M. E. (2016). How intonations interact with tones in Embosi (Bantu C25), a two-tone language without downdrift. In L. J. Downing & A. Rialland (Eds), *Intonation in African tone languages* (pp. 195–222). De Gruyter. https://doi.org/10.1515/9783110503524-007

Roelofs, A. (2015). Modeling of phonological encoding in spoken word production: From Germanic languages to Mandarin Chinese and Japanese. *Japanese Psychological Research*, *57*(1), 22–37. https://doi.org/10.1111/jpr.12050

Savino, M. (2012). The intonation of polar questions in Italian: Where is the rise? *Journal of the International Phonetic Association*, *42*(1), 23–48. https://doi.org/10.1017/S002510031100048X

Schachter, P., & Fromkin, V. (1968). *A phonology of Akan: Akuapem, Asante, Fante*. University of California. https://catalog.hathitrust.org/Record/003051537

Schnur, T. T., Costa, A., & Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of Psycholinguistic Research*, *35*(2), 189–213. https://doi.org/10.1007/s10936-005-9011-6

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. https://doi.org/10.1016/0749-596X(90)90011-N

Shang, P., Roseano, P., & Elvira-García, W. (2024). Cross-linguistic perception of Spanish intonation by Chinese speakers: Effects of linguistic experience and prosodic features. *Porta Linguarum Revista Interuniversitaria de Didáctica de Las Lenguas Extranjeras*, *42*, 127–145. https://doi.org/10.30827/portalin.vi42.27042

Shen, X. S. (1989). Interplay of the four citation tones and intonation in Mandarin Chinese. *Journal of Chinese Linguistics*, *17*(1), 61–74.

Shen, X. S. (1990). *The prosody of Mandarin Chinese*. University of California Press.

Shen, X. S. (1991). Question intonation in natural speech: A study of Changsha Chinese. *Journal of the International Phonetic Association*, *21*(1), 19–28. https://doi.org/10.1017/S0025100300005971

Silverman, K. E. A., & Pierrehumbert, J. B. (1990). The timing of prenuclear high accents in English. In J. Kingston & M. E. Beckman (Eds), *Papers in laboratory phonology* (1st edn, pp. 72–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.005

Sun, L. (1997). Guanzhong fangyan lveshuo [A brief account of the Guanzhong dialect]. *Dialect*, *2*, 106–124.

Sun, L. (2007). *Xi'an fangyan yanjiu [Research on Xi'an dialect]* (1st edn). Xi'an Chubanshe.

Sun, L., & Fu, L. (2019). A study of the phonetic characteristics and rules of Shaanxi dialect. *Journal of Xianyang Normal University*, *34*(3), 33–39.

Thorsen, N. G. (1980). A study of the perception of sentence intonation—Evidence from Danish. *The Journal of the Acoustical Society of America*, *67*(3), 1014–1030. https://doi.org/10.1121/1.384069

Thorsen, N. G. (1985). Intonation and text in Standard Danish. *The Journal of the Acoustical Society of America*, *77*(3), 1205–1216. https://doi.org/10.1121/1.392187

Thorsen, N. G. (1986). Sentence intonation in textual context—Supplementary data. *The Journal of the Acoustical Society of America*, *80*(4), 1041–1047. https://doi.org/10.1121/1.393845

Trager, G. L., & Smith, H. L. (1951). *An outline of English structure*. Battenburg Press.

Truckenbrodt, H. (2002). Upstep and embedded register levels. *Phonology*, *19*(01), 77–120. https://doi.org/10.1017/S095267570200427X

Wan, I.-P. (2007). On the phonological organization of Mandarin tones. *Lingua*, *117*(10), 1715–1738. https://doi.org/10.1016/j.lingua.2006.10.002

Wan, I.-P., & Jaeger, J. (1998). Speech errors and the representation of tone in Mandarin Chinese. *Phonology*, *15*(3), 417–461. https://doi.org/10.1017/S0952675799003668

Wang, J., & Li, R. (1996). *Xi'an fangyan cidian [Xi'an dialect dictionary]* (1st edn). Jiangsu Jiaoyu Chubanshe.

Wang, T., Liu, J., Lee, Y., & Lee, Y. (2020). The interaction between tone and prosodic focus in Mandarin Chinese. *Language and Linguistics*, *21*(2), 331–350. https://doi.org/10.1075/lali.00063.wan

Wang, W. S.-Y. (1967). Phonological features of tone. *International Journal of American Linguistics*, *33*(2), 93–105. https://doi.org/10.1086/464946

Wang, Y. (1995). Shilun Guanzhong fangyan yuyin bianhua de leixin [A preliminary discussion on the types of phonetic changes in the Guanzhong dialect]. *Journal of Yan'an University(Social Science Edition)*, *2*, 91–94.

Wheeldon, L. R., & Lahiri, A. (2002). The minimal unit of phonological encoding: Prosodic or lexical word. *Cognition*, *85*(2), B31–B41. https://doi.org/10.1016/S0010-0277(02)00103-8

Williams, E. S. (1976). Underlying tone in Margi and Igbo. *Linguistic Inquiry*, *7*(3), 463–484.

Wong, A. W.-K., & Chen, H.-C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1172–1190. https://doi.org/10.1037/a0013000

Woo, N. H. (1969). *Prosody and phonology* [Ph.D., Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/14708

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wurm, S. A., Baumann, T., & Lee, M. W. (1987). *Language atlas of China*. Longman Group.

Wyunhe. (2011). *Map of Sinitic dialect—English version* [Graphic]. File:Map of sinitic dialect (Simplified Chinese).svg. https://commons.wikimedia.org/wiki/File:Map_of_sinitic_languages_full-en.svg#/media/File:Map_of_sinitic_languages_full-en.svg/2

Xing, X. (2007). Chinese dialects in Shaanxi province. *Fangyan*, *4*, 372–381.

Xu, B. R., & Mok, P. (2012). Cross-linguistic perception of intonation by Mandarin and Cantonese listeners. *Speech Prosody 2012*, 99–102. https://doi.org/10.21437/SpeechProsody.2012-28

Xu, B. R., & Mok, P. P. K. (2011). Final rising and global raising in Cantonese intonation. *International Congress of Phonetic Sciences*. https://api.semanticscholar.org/CorpusID:46030611

Xu, P. (2006). Guanzhonghua yu putonghua de yuyin duibi [The phonetic comparison between Guanzhong dialect and Putonghua]. *Xiandai Yuwen*, *8*, 88–89.

Xu, X., Lai, W., Li, Y., Ding, X., & Tao, J. (2018). Analysis and modeling of the intonation of Chinese unmarked questions. *Journal of Tsinghua University (Science & Technology)*, *58*(2), 175–180.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*(1), 55–105. https://doi.org/10.1006/jpho.1999.0086

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3–4), 220–251. https://doi.org/10.1016/j.specom.2005.02.014

Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. *17th International Congress of Phonetic Sciences (ICPhS 2011)*, 152–155. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/SpecialSession/Session6/Xu/Xu.pdf

Xu, Y. (2015). Intonation in Chinese. In W. S.-Y. Wang & C. Sun (Eds), *The Oxford handbook of Chinese linguistics* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199856336.013.0012

Yang, J. (2019). Woguo guojia tongyongyu puji nengli jianshe 70 nian: Huigu yu zhanwang [Seventy years of national common language popularisation capacity building in China: Review and prospects]. *Journal of Yunnan Normal University (Philosophy and Social Sciences Edition)*, *51*(5), 41–47.

Yang, Y., Gussenhoven, C., Reshetnikova, V., & Ven, M. V. D. (2024). Functional and phonetic determinants of categorical perception in two varieties of Chinese. *Speech Prosody 2024*, 642–646. https://doi.org/10.21437/SpeechProsody.2024-130

Yao, M. (2014). *Phonetic changes in Xi'an dialect: A study based on the comparison of the old-style and new-style pronunciation of Xi'an dialect* [Master's thesis]. Shaanxi Normal University.

Yip, M. J. W. (1980). *The tonal phonology of Chinese* [Ph.D., Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/15971

Yip, M. J. W. (1988). The Obligatory Contour Principle and phonological rules: A loss of identity. *Linguistic Inquiry*, *19*(1), 65–100.

Yip, M. J. W. (1989). Contour tones. *Phonology*, *6*(1), 149–174. https://doi.org/10.1017/S095267570000097X

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

Yuan, J. (2004). Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling [Ph.D., Cornell University]. In *ProQuest Dissertations and Theses* (305205747). ProQuest Dissertations & Theses Global; ProQuest One Literature. https://www.proquest.com/dissertations-theses/intonation-mandarin-chinese-acoustics-perception/docview/305205747/se-2?accountid=13042

Yuan, J. (2006). Mechanisms of question intonation in Mandarin. In Q. Huo, B. Ma, E.-S. Chng, & H. Li (Eds), *Chinese spoken language processing* (Vol. 4274, pp. 19–30). Springer Berlin Heidelberg. https://doi.org/10.1007/11939993_7

Yuan, J. (2011). Perception of intonation in Mandarin Chinese. *The Journal of the Acoustical Society of America*, *130*(6), 4063–4069. https://doi.org/10.1121/1.3651818

Yuen, I. (2007). Declination and tone perception in Cantonese. In C. Gussenhoven & T. Riad (Eds), *Tones and tunes: Experimental studies in word and sentence prosody* (Vol. 2, pp. 63–78). Mouton de Gruyter. https://doi.org/10.1515/9783110207576.1.63

Zee, E., & Lee, W.-S. (2001). An acoustical analysis of the vowels in Beijing Mandarin. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 643–646. https://doi.org/10.21437/Eurospeech.2001-169

Zhang, C. (2018). Tianjin Mandarin tones and tunes [Ph.D., University of Oxford (United Kingdom)]. In *PQDT - UK & Ireland* (2522361574). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/tianjin-mandarin-tones-tunes/docview/2522361574/se-2?accountid=13042

Zhang, H., Jiang, Q., & Bu, Y. (2017). Putonghua de puji dui Xi'an fangyan de yingxiang—1980 nian zhi jin [The influence of Mandarin popularisation on the Xi'an dialect: From 1980 to the present]. *Anhui Wenxue*, *8*, 158–160.

Zhang, J. Y. (2009). The comparative analysis of the monosyllabic tone between Xi'an and Beijing. *Journal of Shaanxi Institute of Education*, *25*(2), 71–75.

Zhang, J. W. (2018). A comparison of tone normalization methods for language variation research. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Zhang, J. Y., & Shi, F. (2009). The statistical analysis on monosyllabic tone of Xi'an dialect. *Journal of Xianyang Normal University*, *5*.

Zhang, W. (2002). *Yanhua yu jingzheng: Guanzhong fangyan yinyun jiegou de bianqian [Evolution and competition: Changes in the phonological structure of the Guanzhong dialect]*. Shaanxi People's Publishing House.

Zhang, X., Wang, B., Wu, Q., & Xu, Y. (2012). Prosodic realization of focus in statement and question in Tibetan (Lhasa dialect). *Interspeech 2012*, 667–670. https://doi.org/10.21437/Interspeech.2012-205

Zhao, L. (2019). Formation of the interrogative adverb "deisi (得是)" in Xi'an dialect. *Journal of Baoji University of Arts and Sciences (Social Sciences)*, *39*(3), 73–78.

Zhao, Y. (2017). Chinese tone sandhi and the regional differences of citation tones of Shaanxi dialect. *Foreign Language and Literature Research*, *3*(4), 2–13.

Zheng, H. (2004). *Shaanxi Pucheng fangyan yanjiu [A study of the Pucheng dialect of Shaanxi]* [Master's thesis]. Shaanxi Normal University.

Appendix A Statistical Results

Table A.1 Results of acoustic measurements at the phrasal level (Production Study 2)

Phrase	Tone	Pre-target	Tune	Mean intensity	Duration	Mean F0	Max F0	Min F0	F0 range
Pre-focus	T1T1	T1	Question	81.24 (1.72)	418.54 (116.01)	214.62 (57.24)	261.43 (63.41)	170.39 (51.29)	91.04 (36.21)
			Statement	79.58 (1.83)	447.24 (120.03)	188.73 (52.57)	238.18 (75.37)	138.49 (39.55)	99.69 (62.71)
		T2	Question	81.29 (1.79)	443.71 (113.03)	231.09 (58.28)	268.69 (72.60)	190.94 (54.48)	77.75 (58.86)
			Statement	80.45 (1.73)	464.61 (105.84)	199.43 (52.22)	239.02 (77.98)	154.65 (40.76)	84.38 (73.16)
		T3	Question	81.64 (1.27)	452.80 (123.50)	236.36 (58.13)	277.28 (71.07)	185.20 (58.13)	92.08 (55.27)
			Statement	80.74 (1.44)	469.16 (113.98)	204.50 (49.49)	246.30 (64.84)	148.97 (40.83)	97.34 (52.32)
		T4	Question	81.48 (1.55)	433.20 (104.68)	243.52 (59.43)	273.03 (66.26)	205.27 (57.16)	67.77 (45.23)
			Statement	80.39 (1.60)	472.91 (117.81)	209.49 (54.95)	234.07 (60.61)	161.46 (43.89)	72.61 (39.11)
	T2T2	T1	Question	80.97 (1.96)	402.10 (108.80)	211.45 (55.75)	258.89 (64.87)	166.75 (47.78)	92.15 (36.21)
			Statement	79.40 (1.75)	448.82 (120.72)	191.61 (54.17)	248.94 (89.78)	136.30 (38.21)	112.64 (84.63)
		T2	Question	81.11 (1.73)	433.83 (118.33)	231.01 (57.48)	277.92 (85.25)	188.97 (53.95)	88.95 (83.19)
			Statement	80.58 (1.54)	448.06 (102.57)	195.69 (50.29)	232.31 (71.41)	154.87 (41.10)	77.44 (63.94)
		T3	Question	81.48 (1.34)	429.30 (114.67)	232.04 (57.55)	270.74 (65.66)	182.78 (57.96)	87.96 (41.94)
			Statement	80.76 (1.40)	463.61 (120.74)	199.23 (49.77)	240.19 (64.08)	145.44 (39.84)	94.75 (48.26)
		T4	Question	81.09 (1.62)	413.26 (104.14)	239.74 (59.59)	268.33 (70.06)	202.66 (59.95)	65.67 (53.43)
			Statement	80.23 (1.75)	452.82 (108.42)	207.76 (52.64)	232.00 (58.47)	162.55 (43.89)	69.45 (37.00)
	T4T4	T1	Question	80.91 (2.04)	404.26 (119.14)	217.15 (56.64)	269.67 (73.22)	174.42 (51.45)	95.25 (58.40)
			Statement	79.42 (1.61)	436.46 (110.76)	190.79 (51.60)	243.79 (82.83)	140.10 (38.98)	103.68 (73.8)
		T2	Question	81.22 (1.67)	428.63 (109.38)	227.25 (56.41)	272.77 (78.15)	189.78 (50.13)	82.99 (62.35)
			Statement	80.49 (1.60)	456.78 (107.86)	198.96 (51.40)	241.34 (83.52)	153.80 (44.85)	87.54 (78.46)
		T3	Question	81.43 (1.47)	426.65 (115.67)	233.93 (62.72)	271.06 (70.48)	191.06 (65.16)	80.00 (48.63)
			Statement	80.64 (1.43)	464.03 (110.85)	204.17 (49.99)	242.43 (64.47)	153.23 (40.69)	89.2 (52.38)
		T4	Question	81.36 (1.57)	420.07 (110.38)	242.41 (60.40)	270.05 (70.09)	207.28 (57.79)	62.77 (50.32)
			Statement	80.28 (1.73)	458.26 (111.78)	210.97 (50.24)	237.27 (64.59)	163.02 (44.14)	74.26 (54.13)
On-focus	T1T1	T1	Question	83.51 (1.43)	396.61 (97.94)	233.14 (70.44)	283.08 (88.50)	173.28 (51.47)	109.8 (58.51)

			Statement	82.46 (1.95)	424.77 (93.91)	198.67 (55.09)	252.20 (77.89)	134.77 (37.87)	117.43 (65.37)
		T2	Question	83.71 (1.34)	409.88 (101.32)	238.48 (67.29)	288.46 (83.77)	185.97 (56.96)	102.49 (52.59)
			Statement	82.88 (1.78)	437.54 (99.99)	199.97 (57.05)	252.52 (79.68)	140.73 (40.04)	111.79 (61.12)
		T3	Question	83.20 (1.32)	399.84 (98.77)	230.88 (70.25)	279.05 (88.82)	175.97 (52.03)	103.09 (61.29)
			Statement	82.11 (2.27)	427.79 (97.86)	195.01 (56.11)	250.77 (84.47)	136.92 (37.53)	113.85 (71.48)
		T4	Question	83.30 (1.44)	399.64 (97.94)	239.55 (68.34)	287.79 (83.87)	188.06 (55.11)	99.73 (53.86)
			Statement	82.61 (1.63)	421.44 (101.65)	202.59 (56.35)	251.20 (73.03)	143.63 (40.48)	107.56 (50.08)
	T2T2	T1	Question	83.43 (1.40)	414.42 (96.54)	240.92 (69.52)	284.64 (88.51)	172.77 (48.26)	111.87 (56.96)
			Statement	82.81 (1.55)	438.08 (89.84)	195.68 (52.29)	237.31 (67.14)	142.46 (41.59)	94.85 (49.18)
		T2	Question	83.66 (1.41)	420.84 (97.54)	244.36 (69.58)	282.09 (84.4)	200.29 (57.06)	81.79 (51.33)
			Statement	83.26 (1.55)	442.7 (97.78)	198.63 (52.8)	238.17 (71.55)	162.35 (46.1)	75.82 (55.11)
		T3	Question	83.03 (1.40)	408.72 (91.08)	241.09 (70.43)	282.16 (88.32)	182.85 (55.68)	99.30 (58.25)
			Statement	82.51 (1.78)	436.11 (91.92)	194.3 (54.15)	232.89 (67.45)	148.49 (42.68)	84.40 (46.73)
		T4	Question	83.18 (1.39)	416.94 (98.13)	245.44 (69.86)	282.98 (85.32)	198.9 (55.29)	84.08 (48.65)
			Statement	82.9 (1.39)	438.91 (91.41)	201.84 (52.92)	239.04 (67.49)	164.89 (45.38)	74.15 (45.83)
	T4T4	T1	Question	83.81 (1.62)	401.79 (92.17)	275.33 (79.17)	300.19 (88.62)	193.08 (56.3)	107.12 (54.17)
			Statement	83.27 (1.67)	427.26 (90.88)	235.81 (65.63)	261.43 (77.02)	158.27 (45.17)	103.16 (62.03)
		T2	Question	84.19 (1.47)	412.39 (93.61)	279.08 (77.16)	298.21 (89.49)	234.51 (61.58)	63.70 (55.04)
			Statement	83.59 (1.65)	437.65 (98.2)	242.43 (67.72)	263.21 (79.88)	195.62 (50.82)	67.60 (52.59)
		T3	Question	83.47 (1.59)	403.33 (89.49)	272.47 (79.05)	295.84 (86.8)	211.59 (65.78)	84.25 (60.64)
			Statement	83.07 (1.78)	431.8 (93.37)	235.51 (62.71)	260.79 (74.92)	171.87 (48.3)	88.92 (61.41)
		T4	Question	84.02 (1.53)	406.46 (93.55)	283.58 (80.59)	300.83 (91.4)	242.9 (60.98)	57.93 (51.73)
			Statement	83.33 (1.74)	434.42 (105.17)	245.21 (63.46)	264.37 (75.26)	206.2 (49.84)	58.17 (51.59)
Post-focus	T1T1	T1	Question	80.67 (1.53)	567.33 (96.65)	229.53 (57.33)	267.33 (74.21)	188.62 (51.95)	78.72 (52.18)
			Statement	74.83 (3.01)	547.1 (105.44)	145.97 (44.42)	186.33 (72.89)	112.29 (38.42)	74.03 (66.38)
		T2	Question	80.69 (1.63)	571.26 (97.51)	235.27 (59.26)	278.62 (84.88)	191.71 (55.02)	86.91 (63.09)
			Statement	74.83 (2.95)	551.62 (111.33)	145.53 (40.63)	186.62 (70.69)	112.95 (34.38)	73.67 (70.33)
		T3	Question	80.76 (1.51)	563.26 (98.44)	233.8 (58.31)	273.00 (77.59)	190.93 (53.64)	82.07 (55.66)
			Statement	74.75 (3.02)	542.55 (108.44)	144.31 (42.87)	182.73 (73.89)	110.9 (35.11)	71.84 (72.78)

		T4	Question	80.64 (1.58)	565.1 (100.26)	235.04 (58.8)	273.64 (79.6)	194.25 (53.3)	79.38 (58.01)
			Statement	74.86 (3.00)	544.41 (111.74)	144.70 (38.44)	187.06 (74.3)	114.31 (34.28)	72.74 (71.18)
	T2T2	T1	Question	80.91 (1.54)	566.12 (95.34)	247.06 (63.28)	301.04 (85.31)	196.81 (53.84)	104.24 (55.07)
			Statement	77.56 (2.12)	542.37 (111.07)	180.9 (47.17)	248.2 (74.84)	117.72 (38.06)	130.49 (73.42)
		T2	Question	80.79 (1.49)	572.52 (100.08)	245.81 (64.61)	298.88 (83.69)	193.57 (56.02)	105.32 (54.77)
			Statement	77.60 (2.02)	548.57 (108.93)	181.69 (49.96)	250.11 (82.52)	116.27 (38.45)	133.84 (80.37)
		T3	Question	80.93 (1.48)	563.13 (97.6)	246.26 (62.35)	299.85 (85.58)	196.04 (51.99)	103.81 (58.43)
			Statement	77.61 (2.13)	540.54 (111.05)	180.92 (49.25)	247.70 (82.70)	120.05 (39.67)	127.65 (76.99)
		T4	Question	80.88 (1.51)	568.76 (98.21)	248.4 (62.27)	302.62 (83.00)	196.26 (54.55)	106.35 (58.13)
			Statement	77.60 (2.11)	550.25 (108.76)	185.55 (50.1)	253.37 (80.24)	119.53 (38.68)	133.84 (76.85)
	T4T4	T1	Question	80.98 (1.62)	563.27 (94.8)	247.41 (59.08)	309.06 (85.43)	194.39 (49.98)	114.67 (65.47)
			Statement	77.43 (2.01)	540.98 (109.52)	181.91 (49.5)	263.18 (81.57)	118.93 (37.8)	144.25 (81.38)
		T2	Question	80.94 (1.53)	560.73 (90.18)	246.73 (59.13)	307.67 (81.62)	193.98 (50.39)	113.69 (61.8)
			Statement	77.33 (2.07)	543.42 (106.73)	182.2 (50.92)	263.76 (87.16)	118.44 (37.29)	145.32 (85.5)
		T3	Question	80.77 (1.64)	560.06 (99.85)	243.81 (61.23)	306.32 (88.03)	192.12 (54.21)	114.19 (73.66)
			Statement	77.34 (2.01)	537.32 (106.59)	178.8 (45.92)	257.74 (80.74)	120.26 (36.82)	137.48 (81.32)
		T4	Question	80.88 (1.55)	563.25 (93.98)	248.63 (60.94)	311.79 (89.14)	193.46 (52.64)	118.33 (71.46)
			Statement	77.41 (2.00)	541.17 (104.13)	176.96 (45.00)	255.50 (76.98)	117.79 (36.90)	137.71 (73.97)

Table A.2 Results of acoustic measurements of disyllabic words of on-focus phrases (Production Study 2)

Tone	**Pre-target**	**Syllable**	**Tune**	**Mean intensity**	**Duration**	**Mean F0**	**Max F0**	**Min F0**	**F0 range**
T1T1	T1	Syllable 1	Question	83.43 (1.8)	215.59 (58.75)	222.92 (66.63)	276.46 (87.48)	176.62 (51.77)	99.85 (57.21)
			Statement	82.65 (2.28)	234.58 (58.59)	189.11 (55.76)	243.59 (75.99)	145.64 (42.05)	97.95 (60.08)
		Syllable 2	Question	83.29 (1.86)	181.02 (51.28)	240.63 (77.53)	274.98 (87.62)	204.12 (67.64)	70.86 (42.66)
			Statement	81.76 (2.51)	190.2 (49.24)	205.9 (63.31)	242.25 (77.41)	147.17 (43.51)	95.07 (55.84)
	T2	Syllable 1	Question	84.04 (1.52)	223.02 (60.06)	235.41 (63.11)	282.12 (80.54)	198.38 (57.9)	83.74 (48.63)
			Statement	83.3 (2)	245.35 (65.25)	195.69 (55.08)	241.71 (72.98)	160.99 (45.53)	80.72 (52.29)
		Syllable 2	Question	83.05 (1.85)	186.85 (52.78)	239.67 (76.34)	274.61 (87.73)	203 (67.05)	71.61 (47.1)

			Statement	81.84 (2.54)	192.19 (48.64)	200.11 (65.88)	236.1 (82.76)	147.14 (46.59)	88.96 (60.32)
	T3	Syllable 1	Question	82.98 (1.68)	217.86 (60.84)	221.72 (65.17)	272.28 (87.03)	181.84 (53.04)	90.43 (58.43)
			Statement	82.2 (2.62)	234.64 (61.76)	187.9 (58.49)	237.19 (77.22)	150.49 (43.81)	86.7 (60.28)
		Syllable 2	Question	83.13 (1.91)	181.98 (52.41)	237.84 (78.72)	271.12 (90.22)	201.87 (68.33)	69.25 (48.07)
			Statement	81.55 (2.81)	193.15 (49.76)	198.78 (64.03)	235.8 (81.91)	142.96 (40.85)	92.84 (60.35)
	T4	Syllable 1	Question	83.3 (1.86)	214.88 (59.92)	234.85 (64.24)	281.37 (82.42)	198.37 (56.26)	83 (50.64)
			Statement	82.92 (1.82)	228.99 (62.13)	197.23 (54.1)	242.69 (68.94)	165.45 (46.05)	77.24 (42.59)
		Syllable 2	Question	82.96 (1.92)	184.76 (50.15)	242.48 (76.48)	276.68 (86.17)	206.18 (66.35)	70.51 (47.3)
			Statement	81.75 (2.57)	192.45 (54.79)	203.53 (65.24)	237.23 (76.91)	148.15 (45.93)	89.08 (50.36)
T2T2	T1	Syllable 1	Question	83.37 (1.61)	219.03 (53.27)	211.45 (60.69)	262.2 (79.09)	172.77 (48.26)	89.43 (46.4)
			Statement	82.36 (2.02)	233.23 (50.39)	173.67 (49.01)	218.7 (61.92)	142.54 (41.71)	76.16 (38.61)
		Syllable 2	Question	83.15 (2.15)	195.38 (56.58)	262.72 (76.89)	283.21 (87.59)	242.38 (71.26)	40.83 (34.3)
			Statement	83.03 (1.81)	204.86 (53.99)	211.88 (57.46)	233.74 (66.68)	189.85 (52.8)	43.89 (32.7)
	T2	Syllable 1	Question	83.84 (1.51)	227.1 (55.92)	228.91 (62.64)	264.68 (75.49)	200.71 (56.45)	63.97 (42.3)
			Statement	83.26 (1.9)	242.31 (55.91)	188.08 (52.9)	222.08 (67.17)	162.81 (45.82)	59.26 (47.05)
		Syllable 2	Question	83.08 (2.2)	193.73 (53.39)	258.36 (77.62)	278.54 (86.39)	238.88 (72.77)	39.66 (32.11)
			Statement	82.94 (2.05)	200.39 (54.85)	207.33 (56.12)	230.99 (67.17)	186.39 (52.53)	44.61 (35.46)
	T3	Syllable 1	Question	82.71 (1.52)	215.22 (51.15)	217.09 (64.24)	260.88 (79.14)	182.91 (55.72)	77.98 (46.56)
			Statement	81.83 (2.44)	235.23 (53.2)	176.65 (50.95)	216.81 (62.84)	149.17 (42.59)	67.64 (38.24)
		Syllable 2	Question	83.08 (2.12)	193.51 (50.21)	260.06 (76.84)	281.62 (88.29)	239.59 (71.79)	42.02 (36.38)
			Statement	82.89 (1.99)	200.89 (52.89)	208.14 (58.57)	230.53 (67.39)	185.99 (53.67)	44.54 (32.12)
	T4	Syllable 1	Question	83.11 (1.55)	222.81 (55.89)	228.44 (63.13)	265.29 (77.7)	199.01 (55.38)	66.27 (39.76)
			Statement	82.68 (1.57)	235.45 (54.23)	189.92 (49.89)	223.32 (61.04)	165.66 (45.04)	57.66 (34.26)
		Syllable 2	Question	82.91 (2.15)	194.14 (54.82)	260.88 (77.41)	281.21 (85.91)	240.95 (72.24)	40.26 (31.23)
			Statement	82.9 (1.88)	203.45 (49.87)	210.73 (56.52)	235.05 (66.9)	189.36 (51.63)	45.69 (33.66)
T4T4	T1	Syllable 1	Question	83.56 (1.75)	204.94 (50.44)	256.64 (72.52)	287.52 (81.69)	193.36 (56.18)	94.16 (46.1)
			Statement	82.92 (1.92)	220.48 (53.6)	219.3 (59.28)	246.64 (67.04)	158.49 (45.3)	88.16 (44.67)
		Syllable 2	Question	83.74 (2.37)	196.85 (54.06)	291.28 (86.32)	298.5 (89.3)	278.28 (81.04)	20.22 (18)
			Statement	83.35 (2.09)	206.79 (50.42)	250.44 (72.59)	259.32 (78.03)	232.51 (65.17)	26.81 (27.39)

	T2	Syllable 1	Question	84.23 (1.72)	212.95 (55.43)	270.16 (72.56)	289.77 (84.32)	235.24 (61.33)	54.54 (49.46)
			Statement	83.4 (1.93)	226.69 (58.08)	233.08 (62.96)	252.62 (71.95)	197.7 (52.06)	54.92 (39.4)
		Syllable 2	Question	83.84 (1.95)	199.44 (50.07)	287.38 (83.86)	293.93 (86.62)	276.22 (80)	17.72 (15.96)
			Statement	83.53 (1.95)	210.96 (52.47)	251.04 (74.21)	259.91 (79.33)	233.86 (68.13)	26.05 (26.01)
	T3	Syllable 1	Question	83 (1.77)	207.08 (53.01)	259.4 (72.95)	285.01 (80.29)	212.58 (67.33)	72.43 (53.73)
			Statement	82.63 (2.03)	222.8 (54.45)	221.22 (58.29)	247.37 (66.32)	173.1 (50.04)	74.27 (48.75)
		Syllable 2	Question	83.65 (2.23)	196.26 (50.28)	285.23 (85.44)	291.9 (88.85)	273.45 (79.69)	18.46 (17.95)
			Statement	83.27 (2.12)	209 (51.88)	246.99 (69.52)	256.27 (75.02)	228.38 (62.54)	27.89 (27.78)
	T4	Syllable 1	Question	83.8 (1.79)	209.31 (55.03)	274.42 (74.92)	291.41 (83.85)	244.04 (62.34)	47.36 (44.17)
			Statement	82.96 (2.04)	224.53 (70.18)	237.6 (59.64)	254.16 (67.16)	210.38 (51.4)	43.78 (35.87)
		Syllable 2	Question	83.97 (1.95)	197.15 (50.77)	292.48 (87.48)	299.76 (91.42)	280.63 (81.59)	19.13 (19.21)
			Statement	83.45 (2)	209.89 (49.91)	251.98 (69.08)	260.79 (74.86)	233.84 (63.27)	26.95 (38.09)