UNIVERSITÀ PER STRANIERI DI SIENA

Dottorato in Linguistica Storica, Linguistica Educativa, Italianistica

ciclo XXXVIII

Interactive Prosodic Encoding of Tone, Focus and Sentence Type in L2 Mandarin: A Phonetic Study on Italian Learners

Tutor:

Prof.ssa Felicia Logozzo

Prof.ssa Anna Di Toro

Prof. Wen Cao

Dottorando:

Davide Francolino

Anno Accademico 2024-2025

Table of Contents

Acknowledgments

This thesis represents the tangible outcome of numerous human and professional exchanges that have inspired and motivated me throughout my academic journey. I am profoundly grateful for the opportunities these years have offered me to share my research in a variety of beautiful places and academic communities around the world.

I warmly thank my supervisors Felicia Logozzo, Anna Di Toro, and Cao Wen 曹文 for their unwavering dedication and encouragement throughout these years. Their guidance made this journey not only possible but truly inspiring. I am also deeply grateful to the four reviewers of this thesis, whose careful reading and insightful comments have substantially contributed to improving the quality of this work.

I extend my sincere gratitude to Chen Yiya 陈轶亚 for her insightful guidance during the analysis phase, and to Liang Lei 梁磊, whose wisdom helped me find my footing at the very start of this adventure.

I am also deeply appreciative of Claudia Crocco, Sergio Conti, Michelina Savino, Hsu Yu-Yin 許又尹, and Xu Yi 许毅, for their invaluable contributions to the experimental design.

A very special thank you to Andrea Scibetta, a dear colleague and friend, whose unwavering belief in me has been a constant source of motivation. His example of living academic life with kindness, humility, and professionalism has been a steady guiding light through the many challenges and pressures of academia.

I feel deeply grateful to Marco Casentini, my "partner in… GAMM", for his consistent friendly support that made this path lighter, yet richer.

A heartfelt thank you to Carlotta Sparvoli, whose endless inspiration and support carried me forward when I needed it most.

Thank you to Hana Třísková for her trust and dedication, which I treasure immensely.

My sincere thanks go to Jan Lorenc 骆恒, Li Xiang 李翔, Li Yanan 李亚男, Ma Xiaotong 马晓桐, Wachinou Lionnel Pyrrhus Sovi-Guidi 莱昂, Wang Jingjia 王景佳, Wang Rui 王瑞, and all the colleagues and professors from Beijing Language and Culture University who generously lent their help along the way.

I would also like to express my gratitude to Charlie Xu for his generous and dedicated support with statistics. Thanks to Ding Yiran 丁怡然, Katrina Li 李可纯, Zhang Qi 张祺, and all my colleagues at Leiden University Centre for Linguistics, who fostered an environment so welcoming that leaving the office almost felt impossible (literally!).

I am deeply grateful to Alessandra Brezzi for her generosity in providing access to facilities at La Sapienza University for the pilot study, and to Chiara Romagnoli and Giovanni Ielapi for their support in facilitating the experiments conducted at Roma Tre University. My sincere thanks also go to the Centro Servizi Audiovisivi e Multimediali of the University for Foreigners of Siena for their invaluable technical assistance during the experiments conducted at the institution.

Thanks to Hu Rong 胡溶, Lee Jo-Ying 李若莹, Li Xin 李鑫, Zhao Di 赵迪, and Zhong Xin 钟昕 for their precious experimental contributions.

Finally, sincere thanks to the anonymous participants in China at BLCU and in Italy at the University for Foreigners of Siena and Roma Tre University, without whom this research would not have been possible.

While I am deeply thankful for all the insightful feedback received, any errors or oversights in this thesis remain my own.

Grazie alla mia famiglia, per avermi sostenuto incondizionatamente in questo percorso, sopportando spesso la mia assenza, restando sempre presente per me, con pazienza e amore.

Conventions and abbreviations

, **, **	Degrees of statistical significance ($p < .05$, $p < .01$, $p < .001$)
CH	Native Mandarin Speakers Group
EMMs	Estimated Marginal Means
F0	Fundamental Frequency
F0_max	Maximum F0_z value
F0_min	Minimum F0_z value
F0_range	Difference between F0_max and F0_min
F0_slope	Linear slope coefficient from F0_z ~ time (Point) model
F0_z	z-score normalized F0 value
GAMM	Generalized Additive Mixed Model
GLMM	Generalized Linear Mixed Model
IT	Italian Learners Group
L2	Second/Foreign Language
MC	Mandarin Chinese
OtherTone	Variable for adjacent (preceding or following) tone
PFC	Post-Focus Compression
PG	Prosodic Group
PPh	Prosodic Phrase
PW	Prosodic Word
T1, T2, T3, T4	Tone 1, 2, 3, 4
TBU	Tone Bearing Unit
Variable1.variable2	Interaction term between Variable1 and Variable2

List of figures

List of tables

1. Introduction

This thesis examines the production of Mandarin lexical tones by intermediate Italian university learners, focusing on how these tones interact with sentence-level prosody within intonational phrases. While Mandarin tones are phonologically specified at the lexical level, their phonetic realization is shaped by discourse-level factors such as sentence type and information structure. As a result, tones cannot be treated as isolated categories but must be integrated into broader prosodic frameworks – a task that poses particular difficulties for learners from non-tonal L1 backgrounds.

For Italian learners, this challenge is particularly acute, since their native language lacks lexical tone, whereas Mandarin Chinese systematically employs pitch modulation to encode both lexical tone and sentential intonation. Mandarin learners must therefore navigate a dual functional load: preserving the phonological identity of tones while simultaneously adapting them to encode pragmatic meanings such as prosodic focus and interrogativity. The extent to which this integration is successful has important consequences for intelligibility, comprehensibility, and foreign accent perception in L2[1] Mandarin speech.

Despite extensive research on L2 tone acquisition at the lexical level, much less is known about how learners produce tones in prosodic contexts, where local tonal targets interact with global prosodic demands. This gap is particularly evident in relation to Italian learners of Mandarin, whose intonational strategies – e.g., global F0 raising and final rises in yes-no questions – may directly conflict with Mandarin norms.

To address this gap, the present study systematically compares Italian learners' tone productions with those of native Mandarin speakers, interpreting the observed deviations within the framework of the L2 Intonation Learning Theory (LILt; Mennen, 2015). LILt identifies four potential dimensions of cross-linguistic influence – systemic, realizational, semantic, and frequency – which together provide a principled lens for assessing how learners' tonal productions diverge from native norms when embedded in intonational phrases.

The project is structured into three interrelated studies. Study 1 (§ 4) establishes a baseline of tone production in isolation, testing both monosyllabic and disyllabic targets. Study 2 (§ 5) examines the encoding of contrastive focus in disyllabic statements embedded in short

[1] In this thesis, the term L2 is used in a broad sense to refer to languages acquired after the first language, typically during late childhood, adolescence, or adulthood, that is, after the native language(s) have been established. It thus extends beyond the stricto sensu definition of a "second language" as one spoken in the country of arrival, encompassing any language learned after the first (Ortega, 2009; Saville-Troike, 2012). Accordingly, in the present discussion, the terms foreign language and second language are treated as synonymous.

dialogues. Study 3 (§ 6) investigates the interactive realization of sentence type (statements vs. echo questions) across two focus conditions in phrase-final position. Studies 2 and 3 draw on the same corpus and are conceived as subset investigations: Study 2 examines focus production, while Study 3 – more complex in scope – analyses the prosodic encoding of sentence type across focus conditions. Moreover, to maintain comparability across experiments, the target phrases employed in Studies 2 and 3 were segmentally identical to those used in Study 1.

By combining fine-grained acoustic analysis with a multidimensional theoretical framework, the thesis aims to clarify whether Italian learners preserve citation-like tonal contours across contexts, or whether they modify them under intonational pressure in ways that align – or misalign – with Mandarin norms. In doing so, it advances theoretical understanding of suprasegmental acquisition and informs pedagogical approaches to L2 Mandarin, highlighting the necessity of integrating tone accuracy with discourse-level prosody.

1.1 Background and motivation

1.1.1 The physiology of rhythm: on the prosodic foundations of speech

Rhythm constitutes a fundamental property of life across biological, cognitive, and linguistic domains. Prior to its emergence as a linguistic phenomenon, rhythmic organization underlies physiological and behavioral cycles, most notably through the circadian rhythm, the internal clock that regulates the daily alternation between sleep and wakefulness (Refinetti, 2016). Yet among all manifestations of rhythmicity, the auditory rhythm occupies a privileged role. As Sacks (2008) observes, from the earliest stages of development, humans exhibit a spontaneous tendency to impose rhythmic patterns on acoustic input, perceiving and producing sound as an inherently structured, temporal phenomenon.

The prenatal environment itself is highly rhythmic. Contrary to early assumptions that regarded the fetus as a passive organism, research has supported the view that the human fetus is both active and perceptually responsive, equipped with innate sensory and motor reflexes that prepare it for interaction before birth (Lecanuet, 1996; Kisilevsky et al., 2003). By the final trimester of pregnancy, the auditory system is already functionally developed, and by birth, the neural pathways of the auditory nerve are myelinated, allowing for rapid transmission of acoustic information (Hepper & Shahidullah, 1994).

As Querleu et al. (1988) observe, the fetus can hear during the final trimester of pregnancy, showing consistent responses to acoustic stimuli from the 28th week onward. Although sounds from outside the womb are attenuated – rarely by more than 30 dB – external conversations

remain audible, with around 30% of phonetic information transmitted but intonation almost perfectly preserved within the amniotic environment. These findings suggest that the mother's voice and its prosodic patterns can be learned by the fetus, indicating the emergence of short-term auditory memory and an incipient musical-prosodic sensitivity even before birth.

During breastfeeding, infants even adjust the rhythm of sucking in synchrony with auditory stimuli, reflecting an early integration of rhythmic perception and motor coordination (Lecanuet, 1996). At this stage, infants cannot yet identify words or semantic meaning, but they are highly sensitive to prosodic cues – pitch, rhythm, and intensity – that encode emotional and communicative intent (Fernald, 1992; Cutler, 2012). The special speech register known as "motherese", "infant-directed speech" (IDS) or "baby-talk", characterized by higher pitch, expanded intonation range, slower tempo, and exaggerated rhythmicity, serves as a melodic scaffold for early language acquisition (Fernald & Kuhl, 1987; Kuhl et al., 1997). Exposure to IDS has been found to engage neural systems underlying emotion regulation within the orbitofrontal cortex, indicating a close interplay between rhythmic prosody, affective attunement, and early cognitive development (Moschetti, 2007).

As Roach (2001, p. 37) notes, rhythm functions as a cognitive and perceptual mechanism that structures the flow of speech, thereby facilitating access to meaning:

> "[It] helps us to find our way through the confusing stream of continuous speech, enabling us to divide speech into words or other units, to signal changes between topic or speaker, and to spot which items in the message are the most important."

More generally, prosody constitutes a fundamental dimension of language, playing a central role in structuring speech and mediating between physiological, cognitive, and linguistic processes. Although the present thesis does not focus on rhythm per se, but rather on the intonational component of prosody, insights from research on rhythm are nonetheless instructive. Rhythm has been shown to organize human experience from prenatal stages through language development, supporting early speech perception and laying the groundwork for higher-level prosodic and linguistic structures. Far from being a mere acoustic embellishment, rhythmic organization underlies speech segmentation, affective interpretation, and temporal coordination between speaker and listener, thereby providing a neurophysiological foundation upon which linguistic prosody is built.

In sum, according to the literature, research on first-language acquisition consistently shows that prosodic features emerge earlier than segmental units, underscoring their primacy in

speech perception and processing. This body of evidence further suggests that prosody plays a significant role in second-language acquisition, as sensitivity to prosodic patterns can support the development of both perceptual and productive abilities, extending beyond strictly segmental aspects of the linguistic system.

1.1.2 Prosodic competence in L2 acquisition: a theoretical overview

Prosody is a core component of speech organization and interpretation: it signals syntactic structure, manages turn-taking, differentiates utterance types (e.g., questions vs. statements), and conveys speakers' attitudes and emotions. Prosodic cues also shape segmental realization and support the planning and perceptual parsing of speech – functions that are indispensable for successful communication (Zhang & Qian, 2020). Conceptually, prosody functions as an umbrella term encompassing intonation, stress, rhythm, and phrasing – phonetic phenomena realized through temporal, dynamic, and pitch-related parameters such as duration, amplitude, and fundamental frequency (Arvaniti, 2020). Crucially, in tonal languages such as Mandarin, the same physical parameters serve both lexical tones and sentence-level prosody. From an operational perspective, fundamental frequency (F0) variations may be theorised as distinct and organized according to individual communicative functions, yet encoded simultaneously (Xu, 2005).

From the perspective of L2 pronunciation pedagogy, the emphasis has shifted away from the elimination of foreign accent toward the promotion of communicative effectiveness: in this context, the earliest CEFR (2001) formulation of phonological competence drew sustained criticism for its deficit stance toward foreign accents, unrealistic attainment targets, and imprecise use of notions such as stress, intonation, pronunciation, accent, and intelligibility (Piccardo & North, 2017). The recent CEFR (2018) updates reposition intelligibility as the principal criterion for progression and accord prosody the prominence it merits – for instance valorizing the use of prosodic cues to signal information status – recognizing this as an essential skill for L2 learners to express nuanced meanings effectively (CEFR, 2018; Piccardo, 2016).

While the pursuit of a fully native-like accent is neither a realistic nor a pedagogically desirable goal in L2 pronunciation, the development of native-like pronunciation strategies remains essential to support learners in producing speech that is both *intelligible* and *comprehensible* (Abercrombie, 1949; Gilbert, 1980; Pennington & Richards, 1986; Crawford, 1987; Morley, 1991; Pica, 1994; as cited in Yang, 2016). Within this context, three key constructs require clarification: 1) *foreign accent* denotes the extent to which a speaker's

pronunciation diverges from native norms, thereby making their speech sound recognizably non-native; 2) *intelligibility* refers to the degree to which a listener can accurately recognize words or utterances produced by a speaker. Although no universally accepted method exists for assessing intelligibility (Derwing & Munro, 2005; Pickering, 2006), it has often been operationalized through orthographic transcription tasks, whereby listeners attempt to identify the words spoken (Lane, 1963; Kirkpatrick et al., 2008); 3) *comprehensibility*, by contrast, concerns the ease with which a listener can grasp the overall meaning of an utterance in context. Unlike intelligibility, comprehensibility does not require accurate recognition of every word but focuses instead on whether the intended message is successfully conveyed. Importantly, these two constructs, while closely related, are not synonymous: in everyday communication, listeners may comprehend an utterance even without recognizing every word, whereas the recognition of every word does not necessarily guarantee understanding if the utterance's contextual meaning lies outside the listener's knowledge (Yang, 2016, p. 128).

A large body of empirical work, focusing mainly on non-tone languages, has examined how the parameters foreign accent, intelligibility, and comprehensibility covary and how segments versus suprasegmentals differentially contribute to listeners' judgments of L2 speech (see Jesney, 2004; Munro & Derwing, 2011, for reviews). It demonstrated that prosodic deviations, in terms of stress, duration, and intonation, highly contribute to L2 foreign accent ratings and significantly influence L2 intelligibility and comprehensibility (Stockwell & Bowen, 1965; Anderson-Hsieh, Johnson, & Koehler, 1992; Magen, 1998; Hahn, 2004; Munro & Derwing, 2006; Nguyen et al., 2008; Zielinski, 2008). Crucially, given that Mandarin Chinese entail both lexical prosody and phrase-/sentence-level prosody, it is reasonable to expect that both lexical tones and phrase-level prosody in L2 Mandarin Chinese would have a comparable, if not greater, impact on the above-mentioned parameters.

Indeed, according to Yang's (2016) investigation on L2 Mandarin by American learners, prosodic deviations appeared to have even greater impact on both comprehensibility and foreign accent perception, likely due to the vital role of tones in communication. Specifically, the author argues that prosodic deviations from native productions influenced foreign accent ratings by hindering native listeners' comprehension of the target sentences. The study revealed that native Mandarin listeners were highly sensitive to prosodic deviations in L2 Mandarin speech, with comprehensibility scores revealing a strong correlation with foreign accent ratings. This relationship can be attributed to the central role of tone and prosodic accuracy in Mandarin comprehension and underscores the need to address foreign accent, intelligibility, and comprehensibility in an integrated manner in order to promote communicative effectiveness.

Intonation is widely regarded as especially susceptible to cross-language influence in L2 acquisition (Mackey, 2000). By contrast, research on L2 speech has overwhelmingly targeted segments, yielding well-specified accounts of how native and non-native productions diverge and inspiring influential learning models – such as the Speech Learning Model (SLM; Flege, 1995) and the Perceptual Assimilation Model (PAM/PAM-L2; Best, 1995; Best & Tyler, 2007) – whose predictions are anchored in segmental comparisons. The relative neglect of intonation partly reflects its complexity: it interfaces not only with tonal categories but also with tempo, duration, pausing, loudness, and voice quality (Nolan, 2006). This multidimensional nature makes intonation more difficult to delimit, measure, and manipulate experimentally than segmental features.

Research on L2 prosody has attracted increasing attention in recent years, with studies on intonational production expanding the range of target languages under investigation (Avesani et al. 2015). A consistent finding across this body of work is that learners' native language influences their intonational patterns not only at the initial stages of acquisition, but also at intermediate and even advanced levels of proficiency (Mennen, 2004, 2007).

1.1.3 L2 Intonation Learning Theory

Over the last four decades, the autosegmental-metrical (AM) framework has provided a common representational currency for cross-language intonation comparison (Pierrehumbert, 1980; see also Jun, 2005; Ladd, 2008), especially among non-tonal languages. AM distinguishes a finite set of phonological categories (e.g., H [High], L [Low] tones and their combinations) from their continuous phonetic realizations, a distinction that has proven critical for assessing interlanguage patterns (Mennen, 2004, 2007). Specifically, within the AM framework, two fundamental types of tonal events are distinguished: pitch accents and boundary tones.

Pitch accents are phonological cues that associate with metrically strong syllables (typically stressed syllables) and serve to mark prominence within the prosodic phrase. These are notated with an asterisk to indicate their association with stressed syllables, such as H* (high accent), L* (low accent), L+H* (rising peak accent), and L*+H (scooped accent). The nuclear accent, which is the last and most important pitch accent in a phrase, carries particular significance in determining the overall meaning and pragmatic function of the utterance (see Savino, 2012; *inter alia*).

Boundary tones, by contrast, are tones that associate with the edges of prosodic constituents and serve primarily a delimitative function, marking phrasal boundaries. These are notated with the percent sign, such as H% (high boundary tone) and L% (low boundary tone). Additionally, many AM analyses posit an intermediate category called phrase accents, notated with a hyphen (-), such as H- or L-, which represent tones occurring between the final pitch accent and the boundary tone, signaling pitch transitions toward phrase edges. For example, a common intonation pattern at the end of a yes-no question in English might be transcribed as H* L-H%, indicating a high pitch accent followed by a low phrase accent and a high boundary tone, resulting in a characteristic rising contour (Arvaniti, 2022). The ToBI (Tones and Break Indices) annotation system, developed by Silverman et al. (1992), provides a standardized framework for transcribing these tonal events (Beckman et al., 2005).

Building on this framework, the L2 Intonation Learning Theory (LILt; Mennen, 2015) proposes that L2-L1 divergence can arise along four dimensions: (i) systemic (inventory and distribution of categories), (ii) realizational (phonetic implementation: alignment, scaling, shape), (iii) semantic (mapping to discourse meanings), and (iv) frequency (how often structures are used). Mennen (ibid.) provides empirical evidence demonstrating that cross-linguistic differences can manifest across all four of these dimensions.

In the systemic domain, learners may not realize pitch accents absent from their L1 inventories: for instance, Italian and Punjabi learners of English did not produce the complex accents HLH or LHL attested in London English (Grabe, 2004). Learners also deviate in how structural elements combine (accent types, phrase tones, boundary tones), revealing restrictions not only in category sets but also in permissible tune structures.

In the realizational domain, deviations are pervasive: e.g., Dutch learners of Greek align prenuclear peaks earlier than native speakers (Mennen, 2004), and timing mismatches are likewise reported for Korean and German learners of English (Trofimovich & Baker, 2006; O'Brien & Gut, 2010). Scaling differences are common – accents are produced too high or too low (Backman, 1979; Willems, 1982; Wennerstrom, 1994; McGory, 1997) – and learners often alter rise slopes or reduce declination (Willems, 1982; Ueyama & Jun, 1998; Jilka, 2000).

Intonational mismatches in turn affect the semantic dimension. In L2 English, Thai, Japanese, and Spanish learners do not reliably use the high pitch accent (H*) to mark new information (Wennerstrom, 1994), even though it is a canonical function in American English (Pierrehumbert & Hirschberg, 1990); similar difficulties have been observed for Chinese (Juffs, 1990) and Zulu learners (Swerts & Zerbian, 2010). Mandarin-speaking learners of English demonstrate non-native realization of contrastive stress, plausibly reflecting transfer from a

system that marks contrast primarily through duration rather than pitch (Wennerstrom, 1998). Style-dependent modulation also lags behind: even highly proficient learners who can reproduce target intonational patterns often fail to vary them natively across speaking styles (Ulbrich, 2008).

Finally, in the frequency dimension, learners may overuse rises where target varieties prefer falls: Dutch learners of English favor rising accents more than falling ones (Willems, 1982), paralleling the distribution in Dutch and contrasting with English (Willems, 1982; Grabe, 2004). Related substitutions of rises and falls in pitch accents and boundary tones recur across many L1-L2 pairings (Jenner, 1976; Adams & Munro, 1978; Backman, 1979; Hewings, 1995; Mennen et al., 2010; O'Brien & Gut, 2010; Santiago-Vargas & Delais-Roussarie, 2012).

Although these dimensions can interact and are not always cleanly separable – e.g., realizational deviations in scaling or alignment may cascade into semantic miscuing of focus – the LILt framework provides a productive first-pass heuristic for characterizing where and how L2 intonation diverges.

Notably, research on L2 intonation has expanded substantially in recent decades, however, to our knowledge no study has yet applied this framework to the acquisition of a tonal language by learners from non-tonal language backgrounds. The case of Italian learners of Mandarin may therefore offer particularly valuable insights, as intonation in Mandarin must be superimposed upon lexical tones – categories that are absent from the learners' L1 (see § 2 for a detailed discussion of Mandarin prosody).

1.1.4 Prosody and Musicality

A growing body of work points to deep structural commonalities between music and linguistic prosody, including hierarchical grouping, metrical organization, and phrase-final lengthening. Reviews of behavioral, computational, and neurocognitive evidence argue that these parallels make transfer across domains plausible, especially for timing and pitch-based cues (e.g., F0 contours, prominence, and boundary tones) (Heffner & Slevc, 2015).

At the neural level, music and speech-prosody processing share partially overlapping circuitry. Patel's OPERA framework proposes that musical training can enhance speech processing when five conditions are met (Overlap, Precision, Emotion, Repetition, Attention), offering a mechanistic account of why and when musical experience benefits the encoding of speech pitch and timing (Patel, 2011). In line with OPERA, neurophysiological studies demonstrate that musicians (and tone-language speakers) exhibit strengthened subcortical and

cortical encoding of pitch patterns relevant to speech, including more robust frequency-following responses and sharper representation of pitch contours. For instance, Wong et al. (2007) pointed to enhanced brainstem encoding of linguistic pitch patterns among musicians, with concomitant perceptual advantages for pitch tracking in speech. Complementarily, Bidelman, Gandour, and Krishnan (2011) reported cross-domain effects of music and language experience on pitch representation in the human auditory brainstem, indicating shared enhancements for speech-relevant pitch in listeners with musical training and/or tone-language backgrounds.

There is evidence that musical sophistication supports native-language prosody, including improved encoding of intonational pitch movements and better segregation of signal from noise during prosodic processing (e.g., Wong et al., 2007, and previous work cited therein).

Benefits of musical experience extend to L2 prosodic perception and attention to suprasegmental cues. In L2 contexts, musicians display advantages for perceiving prosodic variation in pitch and duration, supporting the idea that musical aptitude can facilitate extraction of prosodic structure in a nonnative language (Marques et al., 2007; Sadakata & Sekiyama, 2011). Marques et al. report that musicians are more sensitive to pitch-violation patterns in speech (behaviorally and electrophysiologically), consistent with enhanced contour tracking; Sadakata and Sekiyama (2011) likewise found musician advantages for nonnative prosody categorization along pitch/duration dimensions.

These findings converge with broader demonstrations that music training sharpens temporal and spectral precision in auditory processing, mechanisms directly implicated in the perception and production of intonational contrasts and prominence patterns in speech (reviewed in Heffner & Slevc, 2015; Patel, 2011).

Taken together, the literature supports a principled expectation that musicality predicts aspects of L2 prosodic competence. In particular, because prosodic meaning in many languages (including Mandarin) is carried by coordinated changes in pitch height, pitch slope, and temporal grouping, individuals with higher musical aptitude may be better at integrating multiple suprasegmental cues – e.g., simultaneously tracking sentence-type contours and focus-driven adjustments – and at stabilizing tone shapes under competing intonational demands.

These considerations support conceptualizing musicality as an individual-difference predictor of higher-order prosodic integration in L2 speech, warranting the inclusion of musical-aptitude indices alongside proficiency in modeling tone-intonation interaction (see §3.5.1.2 for the operationalization of the Musicality variable in this experimental work).

1.1.5 Motivation of the study

The present study addresses the abovementioned gap by systematically comparing the productions of Italian learners with those of native Mandarin speakers and interpreting the observed deviations – across systemic, realizational, semantic, and frequency dimensions – through the lens of LILt. In doing so, it sheds light on the specific challenges that arise when sentence-level intonation must co-exist and interact with lexical tone. More broadly, the study extends the scope of L2 Mandarin research beyond the acquisition of tonal categories in isolation, probing instead into higher-level prosodic domains, particularly the ways in which information structure conditions tone realization.

Although the participant sample represents a specific learner population, the present findings offer a principled foundation for developing L2 Mandarin pronunciation pedagogy targeted at learners from non-tonal L1 backgrounds.

Previous research has largely focused on the acquisition of Mandarin tones at the lexical level (White 1981; Chen 1997, 2000; Zhang 2007, 2010; Xin & Zhang 2009; Xu 2019). A common theme emerging from much of this work concerns the order of tone acquisition among learners: most studies report that Tone 2 and Tone 3 are the most challenging and consequently the last to be mastered (see Sun 1998; Zhang 2007; Xin 2011). However, in many of these investigations, the data have been relatively limited – typically consisting of word lists or isolated sentences assessed according to the production of citation forms, with little or no prosodic context. Futhermore, examining tonal realizations in isolation can be misleading, as native speakers' tones in connected speech may diverge from their citation forms due to contextual and coarticulatory effects (see § 2.3).

By contrast, the present study shifts the analytical focus to tone production within prosodic domains, where pitch contours must simultaneously encode both lexical and intonational meanings, offering a more ecologically valid perspective on the interaction between tone and intonation in L2 Mandarin speech. In native Mandarin, this dual encoding is realized within a single acoustic channel (see § 2.5.2), requiring speakers to coordinate local tonal targets with global prosodic patterns. Such coordination is particularly demanding for learners from non-tonal L1s, such as Italian, for whom the integration of tone and intonation represents a novel and complex challenge.

By examining both production accuracy and context-sensitive pitch modulation, the research evaluates whether tonal misalignments in L2 Mandarin stem from lexical miscategorization or from broader deficits in prosodic integration.

From a theoretical standpoint, the project contributes to the literature on suprasegmental acquisition, specifically the tone-intonation interface in L2 phonology. The data offer empirical grounding for re-evaluating existing models of L2 tone instruction to non-tonal L1 learners.

From a pedagogical perspective, the findings have the potential to inform curriculum design, particularly in highlighting the need for prosody-oriented instruction in L2 Mandarin programs. For learners from non-tonal backgrounds, explicit attention to pitch range control, focus marking, and intonation may help mitigate the observed limitations in connected speech.

1.2 Overview of the project

This project comprises three interconnected studies aimed at examining how Italian learners of Mandarin encode higher-level prosodic information – namely contrastive focus and sentence-type intonation (declarative vs. echo question) – within minimal prosodic domains, specifically disyllabic phrases.

The overarching goal is to identify systematic patterns of prosodic misalignment that go beyond lexical tone misproduction, and that may reflect broader challenges in the mapping between tonal targets and intonational structure.

To this end, the dataset includes F0-based acoustic measures extracted from disyllabic target phrases produced by 42 Italian L2 learners of Mandarin, alongside a reference corpus from 10 native Mandarin speakers. These data are analyzed to investigate whether L2 speakers demonstrate appropriate prosodic modulation in tonal production, or whether their tonal realization remains lexically bound and resistant to discourse-level variation.

The project is structured into three empirical studies:

- Study 1 (§ 4) serves as a baseline, assessing tone identification and production in isolated monosyllabic and disyllabic target words;
- Study 2 (§ 5) explores the prosodic encoding of contrastive focus in disyllabic statements embedded in dialogues;
- Study 3 (§ 6) examines how L2 learners realize sentence-type intonation across tone and focus conditions, specifically focusing on declaratives and echo questions.

Taken together, the three studies aim to clarify the extent to which Italian learners are able to integrate tonal and prosodic cues within minimal intonational phrases, and to assess the degree to which their strategies converge with or diverge from native-like prosodic patterns.

The experimental protocol for data acquisition was composed of three sequential phases:

- Pre-test Phase**;**

- Main Task;
- Post-test Questionnaire.

For the pre-test phase, participants first completed an online tone identification task involving both monosyllabic and disyllabic words, presented in isolation via pre-recorded auditory stimuli. This was followed by a general oral proficiency task, adapted from the HSKK (*Hanyu Shuiping Kouyu Kaoshi*) Intermediate exam, targeting broader spoken competence. The results of these two tasks were subsequently combined to construct the Proficiency variable used in the analyses. All online activities were conducted under researcher supervision via video conferencing, ensuring procedural control and providing immediate assistance as needed.

Immediately prior to the main task, participants completed an in-person tone production task in which they were asked to produce a randomized list of isolated Mandarin words and phrases encompassing all tone combinations. This task, administered individually, was designed to establish a baseline for isolated tone production, and the resulting data constituted the dataset for Study 1.

After a short break, participants engaged in the main reading task, conducted in dyads formed by learners from the same academic cohort. Each pair read aloud short dialogues in Mandarin, designed to elicit contrastive focus and sentence-type intonation within disyllabic phrases. The dialogues were presented in simplified Chinese characters with accompanying pinyin and were designed to reflect authentic conversational patterns, while employing syntactic and lexical structures already familiar to all participants. To minimize potential misproductions resulting from first-time exposure to unfamiliar words, participants were allowed to preview each dialogue before reading and to ask for clarification on any aspect as needed, with the exception of intonational realization.

The reading task consisted of four blocks of approximately 30 dialogues each, interspersed with brief rest intervals. The speech data collected during this phase formed the primary dataset and were subsequently subsetted for the analyses reported in Studies 2 and 3.

After the main task, participants completed a detailed post-test battery comprising a questionnaire on linguistic exposure, learning practices, usage patterns, and affective orientations toward language learning, as well as a tone-deafness test and a self-reported musical aptitude measure.

Scores from the latter two instruments were integrated to form the Musicality variable employed in the analyses.

1.3 Thesis roadmap

This dissertation combines experimental phonetics with second language acquisition research to investigate how Italian learners of Mandarin integrate lexical tones with sentence-level prosody. Its structure is as follows.

Chapter 2 provides an overview of prosodic phenomena in Mandarin Chinese, with reference to Italian where relevant. It serves as a bibliographic survey of the state of the art in Mandarin prosody research, emphasizing those studies most relevant for motivating the present work and for framing the discussion of its results.

Chapter 3 outlines the methodological framework, describing the experimental design, data collection procedures, and the construction of key learner-related variables.

Chapter 4, 5, and 6 examine, respectively, tone production in isolation, tone-focus interactions in statements, and the integration of tone with sentence type and focus on phrase-final position.

Chapter 7 provides a general discussion, synthesizing the main findings across the three studies and evaluating their implications from both theoretical and pedagogical perspectives. It also addresses the study's limitations and identifies avenues for future research that may enrich and extend the present findings.

Finally, the Appendix contains the complete set of experimental materials and supplementary information not included in the main body of the dissertation.

In addition, as specified in Chapter 3, a comprehensive HTML report detailing all statistical analyses can be made available upon request.

2. Prosody of Mandarin Chinese

2.1 Prosodic structure of Mandarin Chinese: an overview

Within an utterance, prosodic constituents are hierarchically arranged, delineating the domains where phonological and phonetic suprasegmental processes are defined (Selkirk, 2003). This prosodic hierarchical structure may function as an intermediary between syntax and phonology and is characterized as "an abstract entity, which is associated with a separate component of the grammar and must integrate various types of information to determine the appropriate prosodic shape of a spoken utterance" (Shattuck-Hufnagel & Turk, 1996, p. 196).

The prosodic structure of Mandarin Chinese (MC), similarly to other languages like Italian, can be analyzed in terms of a hierarchical set of constituents, including the syllable, the foot, the prosodic word (PW), the prosodic phrase (PPh), and the breath group (BG) or prosodic group (PG), as reported in Fig. 1 (Chu & Qian, 2001; Tseng et al., 2005).

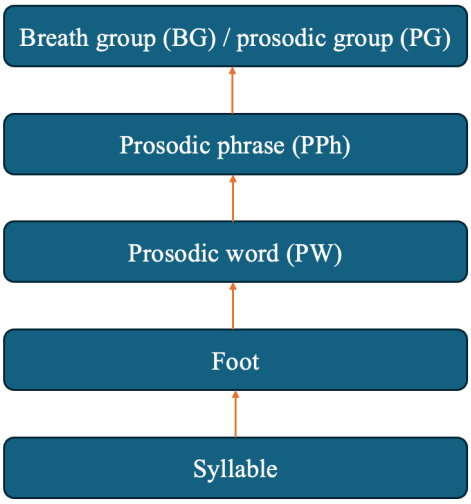


Figure 1 Prosodic hierarchy in MC

2.1.1 The syllable

The syllable is conventionally defined as a cluster of phonemes arranged around a sonority peak, usually coinciding with the vowel, which constitutes the most sonorous and perceptually prominent sound in the sequence (Sorianello, 2006). It is a phonological and phonetic unit on which fundamental prosodic phenomena such as stress and tone are realized.

Its internal hierarchical structure consists of a mandatory nucleus, and optional elements – referred to as the onset (when preceding the nucleus) and the coda (when following it)

(Sorianello, 2006). Broadly speaking, phonemes are distributed along the phonetic continuum in accordance with their intrinsic sonority, determined by the degree of oral aperture during articulation. Vowels, exhibiting maximal sonority, usually form the nucleus of the syllable, around which consonantal segments are organized in progressively lower degrees of sonority relative to their distance from the vocalic nucleus.

The consonant-vowel (CV) syllable type is widely recognized as the most natural and universally attested pattern. Typologically, it occurs in all natural languages and constitutes the first syllable configuration acquired in early language development. Each language, however, is characterized by specific phonotactic restrictions and rules governing syllabification.

In MC, syllables consist of a vocalic or approximant nucleus[2] and, optionally, consonantal elements, and are further characterized by a lexical tone in most morphemes, yet not all (§ 2.2). The structure of MC syllables has been traditionally analyzed using two primary models, which will be discussed in the following subsections.

On the other hand, Italian syllable structure canonically follows a (C)V(C) pattern, with onsets – when present – being relatively simple and predictable, adhering to well-defined phonotactic constraints (Sorianello, 2006; Krämer, 2009; Hermes et al. 2013): the onset may be simple, consisting of any single Italian consonant, or complex, composed of more than one consonant. However, the formation of complex onsets in Italian is not unrestricted. Under the language's syllable well-formedness constraints, the second consonant must be either a liquid /r, l/ (but not /ʎ/), or a glide /j, w/ (Sorianello, 2006). Examples include /**tr**aŋˈkwil.lo/ 'calm', /ˈ**fl**us.so/ 'flow', /ˈ**pj**a.no/ 'flat', and /ˈ**bw**ɔ.no/ 'good'. The coda position is restricted to sonorant consonants (/r, l, n, m/) or to the first portion of a geminate, as illustrated by /ˈpat.to/ 'agreement' and /ˈd͡ʒal.lo/ 'yellow'.

2.1.1.1 Initial-Final model

Traditional Chinese phonologists (Chao, 1968; Cheng, 1973; *inter alia*) argue that tone is a property of the entire syllable, which in turn consists of two segmental components: an initial (I, *shēngmǔ* 声母), which is generally a consonant (C), and a final (F, *yùnmǔ* 韵母), which can comprise up to three constituents, namely a medial (M) or a prenuclear glide (G), a vocalic nucleus (N), and a coda (C), where N is the minimal constituent. The following diagram (Fig.

[2] For a detailed discussion of a proposal in which an approximant functions as the nucleus of a Mandarin syllable, see Lee-Kim (2014).

2) shows the traditional syllable structure of MC according to the Initial-Final model (Třísková, 2011, p. 103):

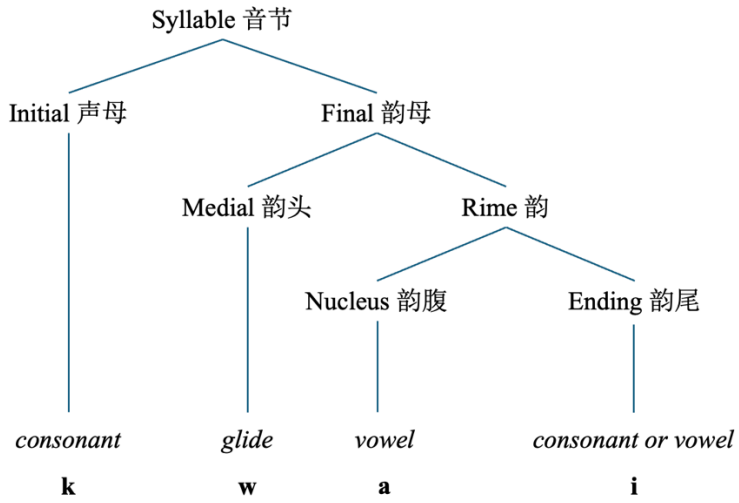


Figure 2 The structure of the syllable 'kuai' 快 according to the Initial-Final model (Adapted from Třísková, 2011, p. 103)

On the Initial-Final model it is based the most common MC romanization system, i.e. Hanyu pinyin, often illustrated in tables with initial and final sounds, as in Fig. 3 below:

	a /a/	**i** /i/	**u** /u/	**ai** /ai/	**ao** /au/	**ei** /ei/	**ou** /ou/
b /p/	*ba*	*bi*	*bu*	*bai*	*bao*	*bei*	
p /p^{h}/	*pa*	*pi*	*pu*	*pai*	*pao*	*pei*	
m /m/	*ma*	*mi*	*mu*	*mai*	*mao*	*mei*	*mou*
f /f/	*fa*		*fu*			*fei*	*fou*

Figure 3 Example table of some MC syllables in Hanyu Pinyin transcription (Adapted from Lin, 2007, p. 283)

The phonological concepts in this model have their origins in ancient rhyming dictionaries, particularly the oldest preserved one, Qièyùn 切韵 (A.D. 601). This dictionary systematically employed a "spelling" system known as *fǎnqiè* 反切, according to which two known characters could indicate the pronunciation of an unknown one, through the Initial sound of the first

character, and the Final sound, including the tone, of the second character[3]. As we might expect, the *fǎnqiè* system was reliable only for a specific variety at a particular historical moment (Handel, 2014, p. 582; Arcodia, Basciano, 2016, p. 93).

As a matter of fact, rhyming dictionaries, such as the Qièyùn, were structured around the concept of the rhyme-carrying subsyllabic component (*yùn* 韵) (see Fig. 2). This component excludes any medial elements (M), allowing syllables with the same rhyme but different M to rhyme. For instance, the modern syllables *dāng*, *xiāng*, and *kuāng* all rhyme according to this concept.

Except for the rhyme component, there are indications that medieval Chinese phonologists had some implicit knowledge of the Medial element and the Ending element; however, the Chinese phonological tradition never explicitly analyzed the syllable into individual segments. Indeed, the further division of *yùn* 韵 into *vocalic nucleus* (*yùnfù* 韵腹) and *terminal* (*yùnwěi* 韵尾) appears to have been introduced under the influence of Western phonological traditions (Třísková, 2011).

2.1.1.2 Onset-Rhyme model

According to the Onset-Rhyme model, syllables are analyzed as hierarchically structured constituents rather than as linear sequences of segments. A syllable consists of an onset (O) and a rhyme (R). The onset may comprise a consonant and, in some analyses, a glide, while the rhyme is further subdivided into a nucleus (N) and an optional coda (Co) (Duanmu, 2007; Lin, 2007; *inter alia*).

Fig. 4 is the diagram of the Onset-Rhyme model for the same exemplificative syllable reported in Fig. 2:

[3] For instance, the character 东 [tuŋ] is described by the *fǎnqiè* formula as 德红反. The first character indicates the onset and the second the rime; accordingly, the pronunciation of 东 [tuŋ] is derived by combining the onset [t] of 德 [tək] with the rhyme [uŋ] of 红 [ɣuŋ], while retaining the same tone as 红 (Wang, 1980).

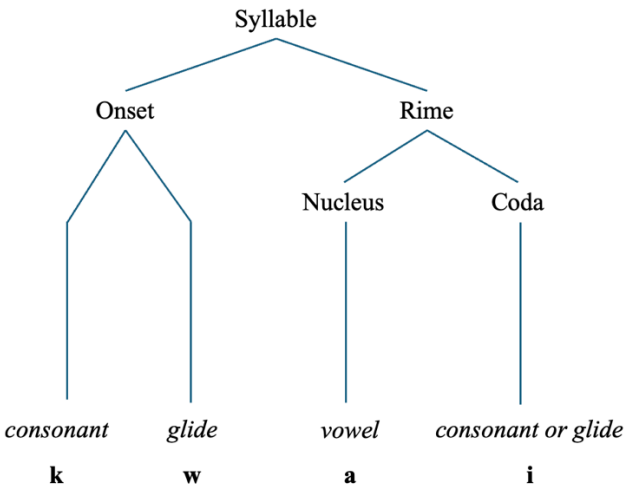


Figure 4 The structure of the syllable 'kuai' 快 according to the Onset-Rhyme model (Adapted from Třísková, 2011, p. 106)

The treatment of the rhyme differs substantially between the Initial-Final model and Onset-Rhyme-based approaches. In the traditional Chinese Initial-Final model, which is non-segmental and non-hierarchical, the syllable is divided into only two components – Initial (声母) and Final (韵母) – without further internal structure. Later pedagogical and textbook versions of this model incorporate a Western segmental perspective, presenting the syllable as a linear sequence of phonemes (often equated with pinyin letters). In this representation, the Final is typically described as comprising a medial (*yùntóu* 韵头), a main vowel (*yùnfù* 韵腹), and a terminal (*yùnwěi* 韵尾), but these components are not hierarchically organized, and the analysis remains largely linear, which may obscure the internal organization of the syllable for learners (Li, 1999; Lin, 2007; Třísková, 2011).

Modern Onset-Rhyme models, by contrast, explicitly impose hierarchical structure on the syllable, though they differ in how closely they align with traditional Chinese categories. Some "traditionalistic" Onset-Rhyme analyses preserve the Initial-Final distinction and further decompose the Final into medial and subfinal components, with the subfinal subdivided into nucleus and terminal, thus maintaining a close correspondence with the traditional notions of *yùntóu* 韵头, *yùnfù* 韵腹, and *yùnwěi* 韵尾; other Onset-Rhyme models, more strongly grounded in Western phonological theory, dispense with the Initial-Final division altogether[4]. In these analyses, the medial (typically realized as a glide) is reassigned to the onset, often

[4] See Třísková (2011) and the references therein for a more comprehensive review of proposed models of MC syllable structure.

forming a complex onset with the consonant, while the rhyme consists solely of the nucleus and an optional coda.

The rhyme component, included in nearly all models of the MC syllable, is described differently in the two above-mentioned models. In the Initial-Final model (Fig. 2) a glide is a part of the final, whereas according to the Onset-Rhyme model (Fig. 4) a glide is part of the onset of a syllable.

Among proponents of the Onset-Rhyme model, there is disagreement regarding the classification of the terminal vowels /i/ and /u/ in syllables such as /mai/, /kʰuai/ and /hau/. Some scholars place /i/ and /u/ within the nucleus of the syllable, resulting in a nucleus with two slots while the coda remains empty. This analysis, which is common in Western phonology, considers falling diphthongs as part of the nucleus. However, this approach disrupts the traditional correspondence between the nucleus and the *yùnfù* 韵腹 (the main vowel), and between the coda and the *yùnwěi* 韵尾 (terminal) (Třísková, 2011, p. 107).

2.1.2 The foot

Unlike the complex tone-bearing units in MC (§ 2.2.1), Italian syllables do not carry lexical tone but may bear stress (§ 2.4). Within the Italian prosodic hierarchy, the foot occupies an intermediate level between the syllable and the phonological word, serving as the domain for stress assignment and rhythmic organization (Nespor & Vogel, 1986; Hayes, 1995).

Cross-linguistically, foot structure is subject to a universal preference for binarity, whereby a foot typically contains either two syllables or a heavy syllable composed of two moras[5] (Hayes, 1995; Vogel, 2009). In Italian, this constraint is reflected in the predominance of disyllabic and trisyllabic feet, often corresponding to paroxytone stress patterns (D'Imperio & Rosenthall, 1999).

Italian feet are generally analyzed as moraic trochees, with primary stress placed on the leftmost mora, yielding a left-headed rhythmic organization (Hayes, 1995). For instance, words such as /ˈka.ne/ 'dog' and /ˈma.no/ 'hand' exhibit a trochaic pattern, with stress on the first syllable. In other, less frequent cases, Italian words may exhibit an iambic foot, that is, a right-headed metrical structure in which the second syllable bears stress. Although this pattern is relatively uncommon, it occurs in certain paroxytonic forms and in prosodic alternations across

[5] A mora is a unit of metrical weight smaller than the syllable, representing the minimal rhythmic timing element in prosodic organization. In many languages, including Italian, a short vowel contributes one mora, while a long vowel or a closed syllable contributes two (Hayes, 1995).

words and phrases. For instance, iambic feet can surface in words such as /tʃitˈta/ ‘city’ or /kafˈfɛ/ ‘coffee’.

In MC, stress is not lexically distinctive, and rhythmic prominence arises instead from the interaction between lexical tone and prosodic phrasing. The comparison highlights a fundamental typological distinction between stress-based and tone-based prosodic systems (Nespor & Vogel, 1986; Hayes, 1995; Duanmu, 2007) (but see § 2.4 for a proposal on the existence of stress in MC).

2.1.3 The prosodic word

The prosodic word (PW) is the smallest unit in speech communication (Wheeldon, 2013) and, in prosodic morphology, may correspond at least to a foot (McCarthy, Prince, 1993). In MC, PWs are typically di- or trisyllabic, though monosyllabic words, such as *tíng* 停 (‘stop’), may also function as a PWs in conversation.

PWs generally begin with a full-toned syllable and serve as rhythmic units, though they don’t always align with lexical words. For instance, the phrase *nǐ měitiān* 你每天 (‘he every day...’) can be articulated as either one or two PWs (Wang, 2003; Yang, 2016). In the former case, the entire phrase constitutes a single prosodic domain with a continuous pitch contour and no internal boundary; the third tone on *nǐ* 你 is therefore likely to undergo tone sandhi before the third tone on *měi* 每 (see § 2.3.1). In the latter case, a prosodic boundary may separate the subject pronoun from the temporal adverbial, and tone sandhi may not apply across this boundary. In fact, the PW is considered the domain for tone sandhi in some Chinese languages (Feng 冯胜利, 1996; Dai, 1998; *inter alia*)

PWs in MC are typically separated by minor breaks without full pauses (Peng et al., 2005; Tseng et al., 2005), and exhibit features such as syllable shortening at the start, pre-boundary lengthening, and pitch discontinuities (Yang, Wang, 2002).

Italian prosodic words typically correspond to a foot structure, often realized as a bi- or trisyllabic layered trochee[6]. Minimal prosodic words in Italian generally exhibit a bimoraic structure which tends to align with both the foot and word edges (Krämer, 2010): this rhythmic organization is evident in processes such as truncation and the formation of hypocoristics (i.e., nicknames), which usually preserve a disyllabic or trisyllabic, trochaic shape. For instance,

[6] A trochee is a metrical unit in which the first syllable is strong (stressed) and the following one is weak (unstressed); for example, /ˈka.ne/ ‘dog’, as cited in § 2.1.2.

Checca (from Francesca) and Fede (from Federico) both conform to this preferred trochaic rhythm, maintaining stress on the first syllable and preserving the bimoraic structure of the word.

2.1.4 The prosodic phrase

The prosodic phrase (PPh) is the next level above the PW in the prosodic hierarchy and typically consists of two or three PWs. The PPh is roughly equivalent to concepts like the intermediate phrase (Beckman, Pierrehumbert, 1986) or the phonological phrase (Nespor, Vogel, 1986). While some researchers suggest that PPh boundaries often align with syntactic constituents (e.g., noun or verb phrases; Nespor, Vogel, 1986; Shattuck-Hufnagel, Turk, 1996; Selkirk, 2003), Ladd (2008) argues that prosodic and syntactic structures should be independently defined.

PPh in MC are typically identified by a noticeable break and a slight pause, along with pre-boundary lengthening and a pitch reset between phrases (Peng et al., 2005; Tseng et al., 2005). The PPh plays a crucial role in the prosodic and rhythmic structuring of many languages, including Chinese (Cao 曹剑芬, 2002; Yang, 2016) and Italian (Grice et al., 2005; *inter alia*).

Italian prosodic phrasing differs fundamentally from MC in both structural organization and boundary marking strategies. In Italian, prosodic boundaries are primarily marked through boundary tones that are phrase-peripheral and associated with specific tonal targets at prosodic edges (Grice et al., 2005; Avesani et al., 2015; see also § 1.1.3).

Notably, previous research has demonstrated that Italian yes-no questions exhibit significant regional variation in boundary tone patterns. Northern and Central Italian varieties frequently employ L-H% rising sequences (contrasting with the terminal fall typical of statements), whereas Southern varieties are characterized by an accentual rise followed by a terminal fall, making the rise in non-terminal position the primary cue for interrogativity (D'Imperio 2002, p. 38). Crucially, Savino (2012) demonstrates that when speech materials are controlled for style (read vs. spontaneous speech), the accentual rise emerges as the predominant feature, and its distribution appears to be independent of geographical variety.

In contrast to Italian, MC prosody relies heavily on pitch reset as the primary and most reliable cue for marking prosodic phrase boundaries (Yang & Wang, 2002; Tseng et al., 2003; Peng et al., 2005). The prosodic system of MC exhibits a clear hierarchical organization, in which prosodic and intonational phrase boundaries are marked by systematic acoustic cues, notably a reset of the lower pitch register and the insertion of pauses. Higher-level prosodic

boundaries are associated with larger pitch resets and longer pause durations, reflecting their greater structural prominence (Yang & Wang, 2002). Crucially, pitch reset in MC is achieved through the bottom line of the intonational contour, creating a systematic declination pattern that resets at boundaries, unlike Italian where boundary marking operates through edge-affiliated tonal targets.

In MC, PPhs at the beginning and end of a prosodic group (PG) exhibit distinct intonation patterns (Tseng et al., 2005). The PG-initial PPhs feature an F0 reset followed by a rapid decline that stops before reaching the F0 minimum (see also declination effect in § 2.5.1). In contrast, the PG-final PPhs also have an F0 reset, though less pronounced than the initial reset, and the contour gradually tapers off with final lengthening; the F0 contours of PG-medial PPhs remain relatively flat, as reported in Fig. 5 below:

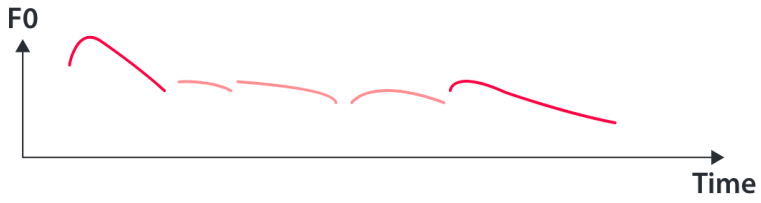


Figure 5 Contours of five PPhs in a PG in MC (from Tseng et al., 2005 and Yang, 2016, p. 20)

In fact, the F0 characteristics of the three PG positions – initial, medial, and final – may reflect distinct intonational roles. In PG-initial PPhs, an F0 reset followed by a non-terminal fall indicates the beginning of a new segment of speech. The flatter F0 pattern in PG-medial PPhs signals continuation, while the lower F0 reset and subsequent gradual decline with final lengthening in PG-final PPhs suggest the approach of a terminal effect.

Tseng et al. (2005) observe both phrase-initial shortening and phrase-final lengthening within PPhs in MC, particularly noting the lengthening of the last two syllables. Cao 曹剑芬 (1999) partially concurs, finding slight initial shortening and final lengthening in PG-initial PPhs, but in PG-final PPhs, she reports initial lengthening and slight final shortening. This suggests complementary duration patterns at the beginning and end of a PG.

The functional differences between Italian and Mandarin are equally striking: while Italian uses boundary tones primarily for discourse coherence and question marking (e.g., with patterns like H% for yes-no questions), MC boundary phenomena mainly serve as hierarchical organizational markers where pause, pre-boundary lengthening, F0 reset and F0 range are

major cues of boundaries with systematically graded acoustic correlates reflecting prosodic hierarchy levels (Tseng, 2002). This contrast reflects the broader typological difference between Italian's edge-based prosodic system and Mandarin's register-reset system for prosodic organization.

2.1.5 The prosodic group

The prosodic group (PG) is the largest prosodic unit in speech characterized by a melodic contour aligning with segmental content and typically marked by pre-boundary lengthening, an audible pause, and pitch reset at boundaries (Beckman & Pierrehumbert, 1986; Himmelmann & Ladd, 2008). In Italian, PGs often correspond to syntactic boundaries but may diverge, especially in embedded sentences. MC, as a tonal and topic-prominent language, presents prosodic contrasts to Italian. Mandarin's prosodic marking at the phrase level is generally less prominent than in non-tonal languages, leading some researchers to conceptualize the PG more as a breath group defined by the limits of one breath (Peng et al., 2005; Tseng et al., 2005).

2.2 Tone

Lexical tone represents one of the most fundamental and complex aspects of phonological organization in a significant portion of the world's languages. In fact, according to Yip (2002) more than sixty percent of the world's languages are tonal, with the majority spoken in Asia, Africa, and Central America.

As a phonological phenomenon, tone refers to the systematic use of pitch variations to distinguish lexical or grammatical meaning, constituting what Hyman (2016, p. 6) characterizes as "any instance where pitch is an exponent of a morpheme in the traditional sense". This definition encompasses a broad spectrum of tonal phenomena, from the relatively simple register tones found in many African languages to the complex contour systems characteristic of East and Southeast Asian languages.

The theoretical understanding of lexical tone has evolved considerably since the early foundational work of Pike (1948) and Welmers (1959), who established the basic parameters for defining tone systems. Modern autosegmental phonology, pioneered by Leben (1973) and Goldsmith (1976), revolutionized our understanding by proposing that tones and their tone-bearing units (TBUs) occupy separate tiers in phonological representation, linked together through association lines (see § 2.2.1.1). This theoretical framework has proven particularly

influential in accounting for the semi-autonomous nature of tone, where tonal melodies can spread across multiple syllables or where single syllables can bear multiple tones to create complex contour patterns.

The functional distinction between lexical tone and grammatical tone represents a crucial theoretical divide in tonal typology. Lexical tone serves primarily to distinguish lexical meanings on monomorphemic roots, as exemplified by the classic Mandarin minimal quadruplet *mā* 妈 'mother', *má* 麻 'hemp', *mǎ* 马 'horse', *mà* 骂 'to scold'. In contrast, grammatical tone functions as a morphological exponent, marking grammatical categories such as tense, aspect, or case. Depending on the language, grammatical tones may attach to toneless morphemes, replace existing lexical tones, or interact with them through tone sandhi. Languages such Thai, Vietnamese and various Sinitic varieties exhibit predominantly lexical tone, whereas languages like Chimwiini display primarily grammatical tone; others, such as Iau, employ both types in complex interactional systems (see Hyman, 2016, for a more comprehensive overview of lexical vs. grammatical tone).

MC represents perhaps the most extensively studied example of a lexical tone language, serving as a paradigmatic case for understanding tone system organization and function. The Mandarin tonal system consists of four primary lexical tones plus a neutral tone, each characterized by distinct pitch contours that are phonologically contrastive.

From a broader theoretical perspective, MC exemplifies what Hyman (2016) characterizes as a language where "tone is largely lexical", contrasting with languages where tone serves primarily grammatical functions. This functional specialization has significant implications for both synchronic phonological organization and diachronic tone system development. The predominantly lexical nature of Mandarin tone correlates with its morphological structure, where each morpheme typically corresponds to a single syllable bearing one of the four contrastive tones.

2.2.1 Tone bearing unit

Tone does not inherently reside within vowels themselves. Instead, it is associated with a phonological unit known as the Tone-Bearing Unit (TBU) (Yip, 2002). The TBU can vary depending on the language and the phonological framework, and may be identified as the syllable, rime, mora, or any sonorant segment (Lin, 2007, p. 92). Proposals regarding the location of the TBU are often motivated by varying theoretical considerations.

Howie's (1974) suggestion of rhyme-alignment is influenced by the significant melodic variability around the syllable onset, attributed to the carryover effects (§ 2.1.1) of tonal influences due to inertia and disruptions caused by initial consonants (Xu, Lee, 2022). The proposal for moraic alignment is primarily motivated by the observation that only relatively long syllables are capable of hosting contour tones, such as rising and falling ones (Duanmu, 1994; Xu, Lee, 2022).

A significant issue concerning the TBU is the accurate definition of syllable boundaries, particularly in relation to anticipatory tonal assimilation (Xu, Lee, 2022). Evidence suggests that the traditional definition of syllable onset – based on the beginning of consonant closure, such as stop closure, frication, or nasal murmur – may be too delayed. Indeed, recent research by Kang & Xu (2024) yielded evidence that tone and vowel articulatory onsets in Mandarin syllables are fully synchronized. This implies that the onsets of consonants, vowels, and tones in Mandarin syllables occur concurrently, and that the revised onset for tone and vowel has shifted the syllable boundary earlier by more than 40% in normalized time.

2.2.2 Fundamental frequency VS pitch

Fundamental frequency (F0) is the primary acoustic correlate of tone and represents the vibration rate of the vocal folds; specifically, it consists of the number of open-close cycles within the vocal folds occurring in one second. F0's measure unit is Hertz, which correspond to one cycle per second.

On the other hand, pitch refers instead to listeners' perception of the F0 signal. Pitch height is directly related to the vibration speed of the vocal folds and, hence, to F0 height.

In terms of articulation, the primary factor determining voice pitch is the tension of the vocal folds (Ladefoged, Johnson, 2010, p. 254). Other phonetic properties, such as the length and thickness of the vocal folds, also condition the production and perception of tone, i.e., men's vocal folds are typically thicker and longer than women's (Kahane, 1978), which is why men's typical mean F0 values are lower than those of women.

Although the two concepts are not identical, they are closely related: higher F0 values generally correspond to higher perceived pitch. In the present study, the terms F0 and pitch are used interchangeably for simplicity, except where explicitly distinguished.

2.2.3 Tone classification and annotation in MC

The four lexical tones of MC are generally described as follows: Tone 1 (hereinafter T1) exhibits a high-level contour, maintained at a steady high pitch throughout the syllable duration;

Tone 2 (T2) demonstrates a rising contour, beginning at mid-pitch and rising to high; Tone 3 (T3) presents the most complex pattern, traditionally described as falling-rising, though phonetic research reveals it is often realized as simply low in connected speech (see § 2.2.4); Tone 4 (T4) exhibits a falling contour, descending sharply from high to low pitch, and is typically shorter in duration than the other tones.

The autosegmental analysis of Mandarin tones, developed by researchers such as Yip (2002) and Duanmu (2007), treats each tone as a combination of high (H) and low (L) tonal features linked to syllabic tone-bearing units. Under this framework, T1 may be analyzed as /H/, T2 as /LH/, T3 as /L/ (with optional final rising to /LH/), and T4 as /HL/.

Chao (1930) proposed a five-level numerical scale to represent relative pitch height in tone languages, in which "1" corresponds to the lowest pitch and "5" to the highest pitch within a speaker's pitch range. This numerical notation is commonly used to transcribe tonal contours by combining digits that indicate pitch height at successive points over the duration of a syllable. Using this system, the four MC tones are conventionally transcribed as 55 (high level), 35 (rising), 214 (dipping or low), and 51 (falling), thus encoding both relative pitch height and contour shape.

The citation forms of the four lexical tones in MC, as represented under the analytical frameworks described above, are illustrated in Fig. 6.

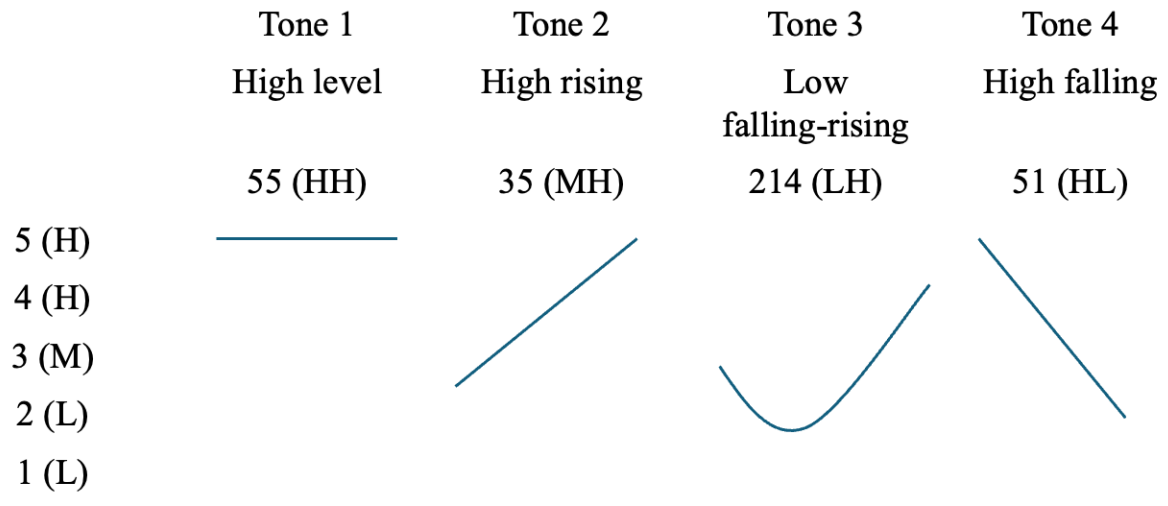


Figure 6 Citation forms of the four lexical tones in MC, adapted from Lin (2007)

One frequently noted limitation of Chao's numerical pitch notation is that, while intuitively clear, it may convey an overly static and idealized representation of Mandarin tones. By

encoding tones as fixed numerical contours (e.g. 55, 35, 214, 51), the system risks obscuring the considerable phonetic flexibility of tonal realization. In practice, even in citation form, tones may surface with different pitch heights or contour shapes (e.g. T1 as 55, 44, or 33; T2 as 45 or 24), and their realization is further modulated in connected speech by factors such as speech rate, prosodic prominence, intonational context, and boundary position. As a result, numerical notation may inadvertently suggest the existence of a single "canonical" contour for each tone, thereby underrepresenting the dynamic, gradient nature of tonal implementation.

In addition to the numerical pitch notation, Chao also introduced the system known as tone letters, designed to represent lexical tones, particularly in languages with contour tones. Developed in the 1920s, Chao's tone letters consist of iconic symbols modeled on a vertical musical staff, created by augmenting the conventions of the International Phonetic Alphabet with a pitch reference scale. Each tone letter represents a relative pitch target or movement within a speaker's pitch range and is typically placed at the end of the syllable it modifies.

Sequences of tone letters can be combined to schematically represent tonal contours, with pitch height mapped onto the vertical dimension of the letter space and terminated by a vertical bar. For instance, the transcription [ma˥] represents the syllable *mā* 妈 'mother' with the first (high-level) tone; [ma˧˥] represents *má* 麻 'hemp' with a rising contour; [ma˨˩˦] denotes the mid-low-rising (dipping) contour of *mǎ* 马 'horse'; and [ma˥˩] corresponds to *mà* 骂 'to scold', characterized by a falling contour.

For the purposes of this study, Mandarin tones are primarily represented using tone category labels (T1-T4) and Hanyu Pinyin with tone diacritics (e.g., *mā*, *má*, *mǎ*, *mà*). Where needed, Chao's numerical pitch notation is additionally employed.

2.2.4 The case of T3 and T4

Numerous researchers demonstrated that the full low-dipping variant of T3 ("214") only occurs in isolation, in focus, and occasionally at utterance final position (Duanmu, 2007; Lin, 2007; Cao 曹文, 2010; Yang, 2016; Wee, 2022). At non-utterance-final position, T3 often surfaces as a low level tone or a half low-dipping tone ("212" or "211") (Chao, 1968).

In fact, T3 classification has been operated differently according to multiple viewpoints. Despite the falling-rising citation form still being used in L2 Mandarin language teaching, some scholars have demonstrated that the falling and rising contours of T3 may be treated as phonetic variations that are not perceptually relevant. Therefore, it has been proposed to classify T3 phonologically as a low-level tone (see Sparvoli, 2017; Li, 2021 and references therein).

According to Chao (1968, pp. 28-29), T4's ending point is higher when it follows another T4. In fact, in connected speech and in non-phrase-final positions T4 often manifests as 52 or 53 (Lin, 2007, p. 96; Yang, 2016; see also § 2.3.1).

2.2.5 Tone production mechanisms

In speech, pitch contour is affected in multiple ways at the suprasegmental level (e.g., according to stress patterns, sentence moods, etc.), but also at the segmental level (e.g., initials like voiceless fricatives and aspirated affricates can raise pitch contour, while the sonorant and lip sounds can lower pitch contour; Li, 2002; Xu, Wang, 1997; *inter alia*).

A tone may frequently "misalign" with its corresponding syllable due to transitional mechanisms that Xu & Wang (1997) call late alignment, dynamic overshoot and anticipatory raising.

Late alignment occurs as the initial segment of the syllable (usually the initial consonant) is still involved in the transition from the preceding syllable, therefore its F0 variation differs according to the preceding tone, while the proper syllable tone is realized mainly in the later portion, gradually converging into the specific tone contour.

As a result of the late alignment, a dynamic overshoot mechanism may occur when a contour tone is realized with a rapid F0 movement through the end of the syllable and in the initial part of the following syllable possibly due to articulatory inertia. Dynamic overshoot also accounts for peak delay, a phenomenon in which the F0 peak expected within a syllable is instead realized in the following one. This phenomenon has been observed in both tone and non-tone languages (Arvaniti, Ladd, 1995; Xu, Wang, 1997 and references therein).

Anticipatory rising occurs when a tone's high F0 region is raised when followed by a low tone. Xu and Wang (1997) claim that late alignment and anticipatory raising are likely both contributing factors to the downstep phenomenon (see § 2.5.1). In fact, as already noted above, a recent study by Kang & Xu (2024) presents compelling evidence for the complete synchrony of articulatory onsets between tone and vowel in Mandarin syllables. This finding supports the synchronization model of the syllable, which posits that consonantal, vocalic, and tonal elements are articulated simultaneously, effectively constraining most temporal degrees of freedom. A significant implication of this research is that the shared onset of tone and vowel shifts the syllable boundary earlier, indicating that the previously noted anticipatory raising of the F0 peak takes place within the syllable carrying the low-pitched tone, rather than before it.

2.3 Tonal processes in connected speech

Within the PW and higher-level domains (see § 2.1), tonal realization may vary in both pitch level and contour. Tone changes can result from tonal contexts and morphosyntactic conditions (i.e., tone sandhi), but also due to prosodic constraints, speaker characteristics, and local factors related to tonal contexts and syllable composition (Xu & Lee, 2022).

Contextual tonal variation has often been discussed as encompassing "tonal coarticulation" phenomena (Shen, 1990b; Xu, 1994); however, as Xu and Lee (2022) suggest, this term is problematic for at least two main reasons. First, coarticulation traditionally refers to the overlap of consonant and vowel gestures within a syllable due to simultaneous movements of different articulators, which doesn't apply to tone since it involves only the larynx. Second, coarticulation suggests that a phonetic unit adopts features from a neighboring unit, typically in an assimilatory manner; however, not all tonal variations are assimilatory; for example, they can include phenomena such as carryover effects, anticipatory effects, and consonantal perturbation, in addition to tone sandhi.

In this section, tonal variation phenomena are examined, starting with the most well-known, though not least problematic: tone sandhi.

2.3.1 Tone sandhi

Tone sandhi refers to the systematic alteration of a tone when it occurs in specific phonological or morphosyntactic contexts. These variations are influenced by the surrounding tonal environment and the global prosodic or morphosyntactic structure in which the tone appears, giving rise to predictable phonological alternations (Chen, 2000; Zhang, 2014).

For instance, in Taiwanese Southern Min, tone sandhi applies within prosodic domains, where a non-final syllable in a phrase changes tone depending on the following syllable's tone (e.g., *sió* 'small' → *siáu* in *sió-kháu* 'small mouth'; Chen, 2000). Similarly, in Shanghai Wu, tonal alternations are largely syntactically conditioned, spreading across an entire phrase rather than within individual words (Zhang, 2014).

Tone sandhi is classified as phonologically motivated, as it follows specific rules dictated by phonological principles. This classification contrasts with other contextual tonal variations, which are generally phonetic processes more variable and influenced by factors such as the speaker's rate and style (Zhang, 2022).

In this subsection, we will examine tone sandhi processes that are widely recognized in the literature as phonological sandhi, as well as phonetic processes, which some scholars refer to as phonetic sandhi.

One of the most studied phenomena of phonological tone sandhi in MC (as well as in Xiamen Southern Min) involve (1) T3 becoming a rising tone when it precedes another T3. This sandhi rule is always included in L2 Mandarin textbooks, although it is often illustrated only within disyllabic words, e.g., *nǐ hǎo* 你好 'hello' is pronounced *ní hǎo*. In fact, the phenomenon becomes more complex in polysyllabic words, as the rule is influenced by syntactic, prosodic, and semantic factors within the tone sandhi domain (see § 2.6.3).

On the other hand, a more phonetically motivated T3 sandhi is what is also called (2) T3 reduction: T3 is realized as a low tone when it precedes a non-T3 (Yang, 2016, p. 8):

老师

Lǎoshī

LL HH

Reduction phenomena are also observed in T4, which exhibits pitch-range compression when followed by another tone, i.e. 51 (or HL) of its citation form becomes 53 (or HM).

The second phonological sandhi to be mentioned concern specifically the words *yī* 一 'one' and *bù* 不 'no'. In non-phrase final position, before a lexical tone different from T4, T1 in *yī* 一 (55, HH) becomes T4 (51, HL), and follow T4 reduction rule, as reported in Fig. 7 below on the phrase *yī wǎn* 一碗 'one bowl':

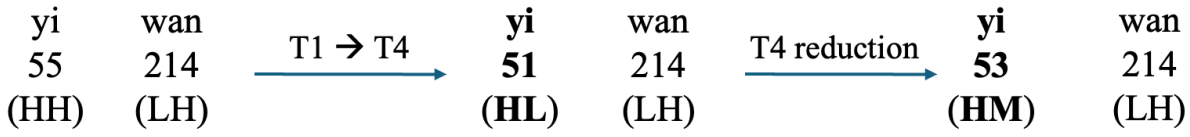


Figure 7 Example of a phonological sandhi in the word *yī* 一 and consequent T4 reduction

When followed by a T4, T1 in *yī* 一 (55, HH) and T4 in *bù* 不 (51, HL) become T2 (35, MH). This phenomenon may be considered as a dissimilation, although it only applies with *yī* 一 (cfr. *yí yàng* 一样 'the same') and *bù* 不 (*bú yào* 不要 'do not want/have to'; Lin, 2007, p. 199).

A further frequently mentioned phonetic sandhi involves T2, which changes to T1 when it follows T1 or T2 and precedes a non-T0, as in *yóu* 油 in *cōng yóu bǐng* 葱油饼 (Chao, 1968; Lin, 2007; Yang, 2016). This type of sandhi is considered phonetic rather than phonological, as it occurs more frequently in casual, fast speech in prosodically weak positions, such as the word-medial syllable in the example above.

2.3.2 Tone target undershoot

A phenomenon similar to T2 phonetic sandhi is tone target undershoot (Lindblom et al., 1990; Yang, 2016; Xu & Prom-on, 2019), which mostly occurs in casual, fast or moderately fast speech. Tone target being the phonological components of tones, target undershoot refers to a phenomenon whereby, in unstressed (i.e., phonetically weak) syllables, these targets are not fully reached due to insufficient time of articulation and weak articulation strength (Xu & Lee, 2022). An example of tone target undershoot can be observed in the second syllable (ming2, i.e. *míng* 明) of the Fig. 8 below, by Yang (2016, p. 55):

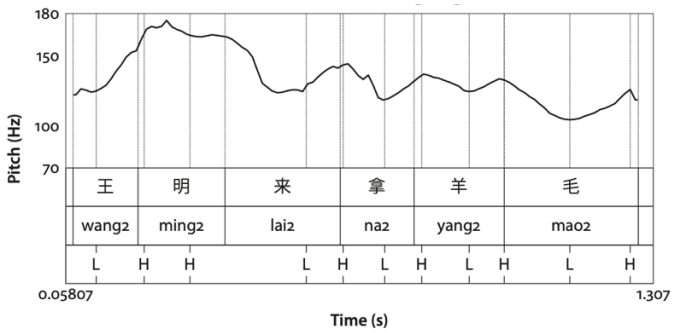


Figure 8 Example of tone target undershoot in a T2 sequence PPh

In extreme, albeit infrequent, instances, target undershoot may be so strong that an entire syllable, including its tonal characteristics, appears to be entirely omitted (Cheng, Xu, 2015). Yang (2016) argues that these phenomena rarely appear in teachers' talk, which may be one of the reasons why students are often unaware of such phonetic tone changes in connected Mandarin speech.

2.3.3 Carryover effects

Carryover effects occur when the production of a tone is influenced by the tonal or articulatory properties of the preceding syllable. In such cases, features of the earlier articulation are retained and carried over into the subsequent sound, thereby shaping its

phonetic realization (Crystal, 2008, p. 82). However, not all carryover effects are assimilatory; some are dissimilatory in nature, leading a tone to diverge from, rather than converge with, the characteristics of the preceding tone.

Carryover assimilation effects can persist throughout an entire syllable or even extend into the subsequent syllable. However, their influence typically diminishes over time and nearly vanishes by the end of the second syllable (Xu, 1997, 1999). This is illustrated in Fig. 9 from Xu and Lee (2022), where the tone of the second syllable (T1 in the example below) remains constant while the tone of the first syllable changes across four different variations.

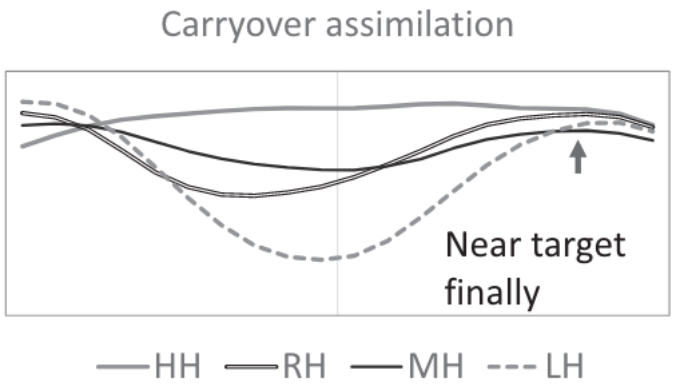


Figure 9 Exemplification of carryover assimilation effect (from Xu and Lee, 2022)

As reported, the F0 trajectories of the second syllable consistently begin from the final F0 of the first syllable and then progress smoothly toward the high-level contour characteristic of the target tone, i.e. T1. The gradual alignment of tones with a supposed pattern led Xu and colleagues to claim that tone articulation is aimed at reaching a stable underlying tonal target (Xu, Wang, 2001; *inter alia*). The extended duration required to achieve a tonal target is attributed to speech rates nearing the maximum speed of pitch change (Xu, Sun, 2002). Consequently, a significant portion of the articulation time is dedicated to transitioning toward the target rather than maintaining it.

The time required to reach the tonal target is influenced by several factors, including the direction of pitch change (ascending or descending), the distance between adjacent tonal targets, and the articulatory strength, or muscular force, involved (Xu, Lee, 2022). This is the reason

why tone target undershoot generally appears in conflicting tone contexts (e.g., T2-T2 combinations) rather than in compatible ones (e.g., T2-T4 combinations)[7].

Carryover effects can also exhibit dissimilatory characteristics. One such effect, known as post-low bouncing (Chen, Xu, 2006; Prom-on et al., 2012), is induced by a tone with a very low pitch and has been observed both in Cantonese and MC. This phenomenon is conditioned by the pitch level of the triggering tone. In MC, post-low bouncing occurs typically when a L tone is followed by one or more T0, as illustrated in the continuous contour in Fig. 10 (with post-low bouncing indicated by the grey arrow), or when the L tone is placed under focus.

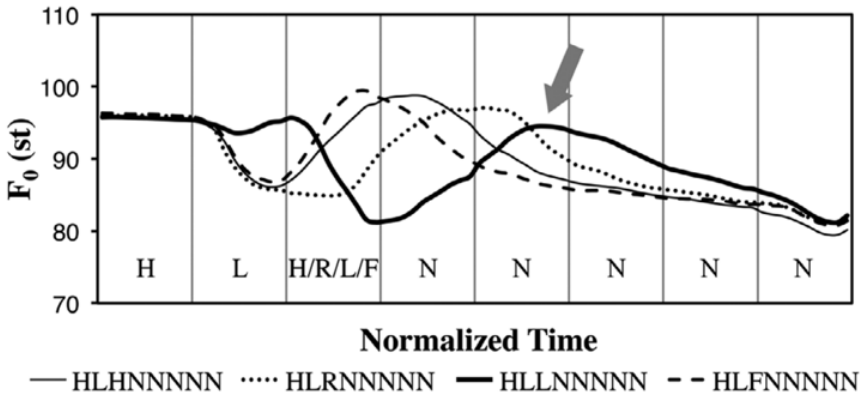


Figure 10 Example of post-low bouncing after a L tone (from Xu, Lee, 2022)

Anticipatory effects involve the impact of a tone on preceding tones, typically affecting only the final portion of the immediately preceding tone. Indeed, these effects are highly localized (Shen, 1990b; Peng, 1997; Xu, Lee, 2022). The most frequently observed phenomenon is pre-low raising, also known as F0 polarization (Hyman, Schuh, 1974), as well as anticipatory dissimilation (Xu, 1997) and anticipatory raising (Connell, Ladd, 1990; Xu, 1999).

Pre-low raising is a local variation where the F0 of a H tone increases before a L tone (Fig. 11). This phenomenon has been clearly documented in several Chinese languages, including Cantonese (Gu, Lee, 2009), Fuzhou Min (Li, 2015), and MC (Xu, 1997). As Xu & Lee (2022) note, a similar phenomenon has been observed in singing (see 'preparation' in Saitou et al., 2005), suggesting that the phenomenon is likely related to articulatory processes.

[7] A congruent (or compatible) tone context occurs when the tone targets of adjacent syllables align, meaning the phonological value at the offset of the preceding syllable matches that at the onset of the following syllable; e.g., a rising tone before a falling tone (T2-T4), where the high offset of the rising tone is congruent with the high onset of the falling tone (Xu, 1994).

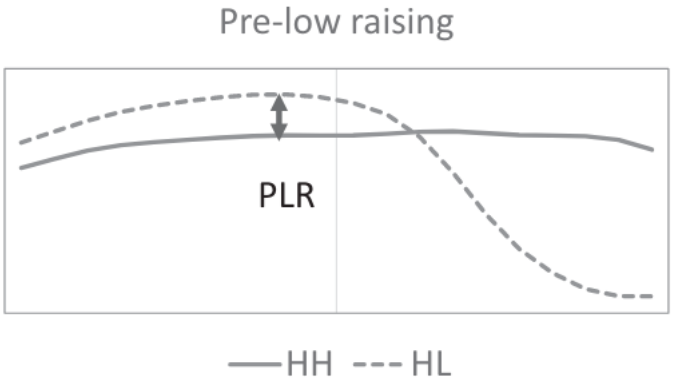


Figure 11 Example of pre-low raising (Xu, Lee, 2022)

2.3.4 Effects of segmental features on tone production

There are two key effects of segmental features on tone production: vowel-intrinsic pitch and consonantal perturbation. Research widely demonstrated that low vowels generally have a lower F0 than high vowels across languages (Whalen, Levitt, 1995), possibly due to laryngeal configuration (Sapir, 1989), subglottal pressure (Steele, 1986), and strategies to enhance vowel contrast (Diehl, Kluender, 1989). The impact of vowel height on F0 is most significant for isolated tones and sustained vowels, but is much less pronounced in connected speech (Ladd, Silverman, 1984). However, even in connected speech, this effect must still be accounted for when carefully comparing F0 across tones (Xu, Lee, 2022).

The second effect is the well-documented consonantal perturbation (Hombert et al., 1979). Typically, voiceless consonants tend to raise the F0 of the following tone. In MC, for example, a voiceless unaspirated stop increases the F0 of the following vowel compared to both voiced and voiceless aspirated stops (Xu, Xu, 2003; Xu, Lee, 2022).

Hombert et al. (1979) popularized the view that voiced consonants significantly lower F0; however, Silverman (1986) demonstrated that when intonation is carefully controlled, both voiced and voiceless consonants actually raise F0 at voice onset. Xu & Xu (2003) further supported this claim, as reported in Fig. 12, where F0 contours of the same tones with different onset consonants align on the final segment, where the tone target is reached. This alignment reveals that voiceless consonants consist of three phases: (a) an unvoiced portion, (b) a brief aerodynamic perturbation, and (c) a final segment reflecting the underlying target.

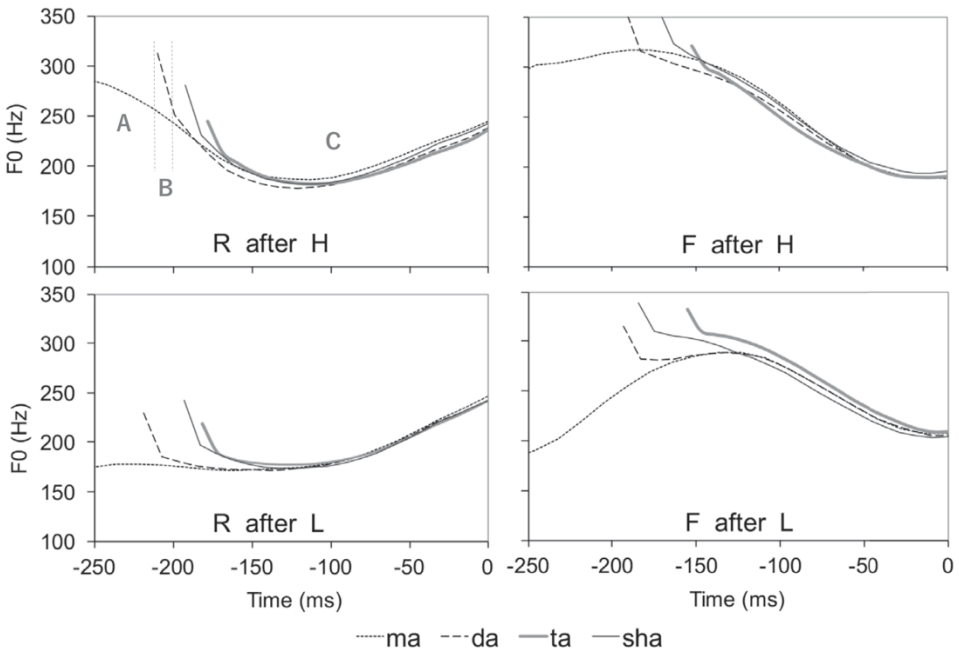


Figure 12 Effects of voiceless consonants on the F0 contours of MC (from Xu, Xu 2003; Xu, Lee, 2022)

Compared to the continuous initial sonorant contour (*ma*), where the initial F0 is influenced by the preceding syllable, the terminal parts (c) of the other contours are largely similar. However, it is important to note that this interpretation depends on how the F0 contours are aligned. Aligning the contours at voice onset would clearly yield different results. Therefore, the chosen alignment must be justified to accurately compare productions with different segmental features.

2.4 Stress in Mandarin Chinese

Several studies have demonstrated, through different methodological approaches, that the human brain encounters significant difficulty in perceiving and retaining sequences of indistinguishable items in the absence of a discernible pattern (Miller, 1956; Bianco et al., 2020; Vetchinnikova et al., 2023).

In speech perception, stimuli with varying levels of prominence are more easily processed compared to a continuous stream of syllables or words with uniform prominence. This variability in prominence underscores the importance of speech rhythm, which plays a pivotal role in facilitating comprehension and effective oral communication (Steele, 1986; see also § 1.1.1).

In speech, rhythm is often understood as the alternation between stressed and unstressed syllables. Similarly to tone, stress is a suprasegmental feature with the syllable or the foot as its domain. Stress is identified on a relative basis: perceptually, a syllable in a word is stressed if it is more prominent than the others. Acoustically, stress is realized in a language-specific manner, typically involving variations in pitch, duration, and/or intensity (Lin, 2007; Eriksson et al., 2016). In terms of articulation, producing a stressed syllable involves greater respiratory energy and increased laryngeal activity (Ladefoged, Johnson, 2010, p. 111).

Although traditionally classified as syllable-timed, Italian displays acoustic characteristics that complicate this categorical distinction (Braun, Geiselmann, 2011). Rather than strictly controlling inter-stress intervals (as in prototypical stress-timed languages like English), Italian achieves rhythmic organization through a coupled system of timing mechanisms: speakers lengthen stressed vowels and modulate pitch and intensity to mark prominence across multiple stressed syllables within longer words (Eriksson et al., 2016). Longer words often feature multiple stressed syllables with varying degrees of prominence, where the most prominent syllable is assigned primary stress, and the others are given secondary stress. The following examples[8] illustrate the realization of primary and secondary stress in Italian, where primary stress is marked by [ˈ] and secondary stress by [ˌ] according to IPA conventions.

[eˌlet.tri.t͡ʃiˈta] ‘electricity’
[enˌt͡ʃe.fa.loˈgram.ma] ‘encephalogram’

The rhythmic nature of stress alternation is phonologically represented through the foot structure. Within each foot, one element is obligatorily more prominent than the other, yielding a recurring rhythmic pattern. This is illustrated in Fig. 13, adapted from Lin (2007), for English, where the first syllable of the second foot (/ˈneɪ.ʃən/ ‘nation’) is more prominent than the first syllable of the first foot (/ˈɪn.tu/ ‘into’):

[8] From Brugnoli (2024, p.43).

in	*to*	*na*	*tion*	
(σ	σ)	(σ	σ)	σ = syllable; ( ) = foot domain
x		x		x indicates prominence/stress
		x		

Figure 13 Stress alternation expressed by stress feet (from Lin, 2007)

Some languages like Italian or English generally feature a left-prominent foot (see § 2.1.2), where the left element of a foot carries more prominence or stress, as represented in Fig. 14a. Conversely, some other languages like French or Turkish have a right-prominent foot, with the right element of a foot being more prominent or stressed[9], as outlined in Fig. 14b. Both systems exhibit stress alternation, but the rhythmic patterns differ: one follows a stressed-unstressed pattern, while the other follows an unstressed-stressed pattern.

a. left-prominent foot	b. right-prominent foot
(σ σ)	(σ σ)
x	x

Figure 14 left- VS right-prominent foot (from Lin, 2007)

The presence and nature of stress in MC remain a topic of debate. As Duanmu (2007) observes, although word-level stress is largely rejected in MC (Wang 王力, 1958; Hyman, 1977; inter alia), the language nonetheless exhibits systematic variations in syllabic prominence. These variations are essential to characterize for a comprehensive understanding of MC speech rhythm. Phonetically, syllabic prominence in MC is typically manifested through expanded pitch range, increased duration, and heightened intensity (Chao, 1968, p. 35; Shen, 1990a).

The clearest and least disputed instance of word-level prominence in MC occurs in disyllabic words containing a neutral tone (see Fig. 15). In these words, the first syllable retains its full lexical tone, whereas the second syllable is toneless and phonologically less prominent, resulting in a left-prominent foot structure (Lin, 2007, p. 224).

[9] Cfr. /paʁ.ləˈmɑ̃/ ‘parliament’ in French or /ki.taˈpɯ/ ‘the book’ in Turkish (Jun & Fougeron, 2002; Kabak & Vogel, 2001)

(3)	*dōngxi*	[tuŋ]55 [ɕi]2	'things, stuffs'
tone:	T1 T0		
foot:	(σ σ)		
stress:	x		

Figure 15 Possible instance of word stress in a disyllabic word featuring T0 in MC (from Lin, 2007)

However, in most MC words, each syllable carries a full lexical tone, and the distribution of prominence is not straightforward, making it difficult to establish consistent stress patterns (Chao, 1968, p. 238; Lin, 2007, p. 224), even though in fast, spontaneous, or casual speech a substantial number of syllables may undergo tonal weakening or neutralization (see § 2.3). This is in stark contrast to Italian and English, where the stressed syllable in a word is easily identifiable.

Duanmu (2007) suggests that the inconsistency in identifying word stress in MC arises because the primary phonetic cue for stress is F0 variation, which is already employed for multiple functions, including distinguishing word meaning through lexical tones and conveying utterance intonation. The author argues that although native Chinese speakers may find it difficult to consistently identify main stress in words and Chinese dictionaries do not annotate stress, there are observable features that differentiate stressed from unstressed syllables. Additionally, other aspects of stress in Chinese, such as phrasal focus, are realized in ways similar to word stress (Duanmu, 2022).

In fact, some form of phonological foot structure may be identified in MC, as there is a strong tendency for Mandarin words to be at least disyllabic (Duanmu, 2000). Indeed, in MC it is common for a monosyllabic morpheme to become disyllabic through reduplication (e.g., *mèi* 妹 'younger sister' becomes *mèimei* 妹妹) or by combining with another often-unstressed morpheme (e.g., *hǔ* 虎 'tiger' becomes *lǎohǔ* 老虎, *xué* 学 'study' becomes *xuéxí* 学习). Country names follow a similar pattern, with a disyllabic form required for monosyllabic names (e.g., *Měiguó* 美国 for 'America', *Hánguó* 韩国 South Korea) by adding *guó* 国, but not on already disyllabic names (cf. *Rìběn* 日本, not **Rìběnguó* 日本国; *Cháoxiǎn* 朝鲜 not **Cháoxiǎnguó* 朝鲜国) (Duanmu, 2007, 2022; Lin, 2007).

On this base, Duanmu (2007, 2022) proposes two major foot structures for disyllabic words: Heavy-Light and Heavy-Heavy, both with initial stress, consistent with Chao (1968). For compounds with three or more syllables, it is claimed that the stress pattern is generally shaped by internal syntax. Chao (1968) instead proposed a relative prominence pattern for words

composed entirely of full syllables, which he characterized as strongest-next-weakest. In Duanmu's (2007, p.142) terms, this corresponds to a 2-x-1 pattern, where "1" denotes the strongest syllable, "2" a less prominent one, and "x" represents one or more syllables that are weaker than "2". Other analyses suggest that in these words, disyllabic feet are formed from left to right (Shih, 1997; Feng, 1998; Chen, 2000; Duanmu, 2007, 2022).

Among the many diverse views on MC stress, Třísková (2019) proposes an alternative perspective, claiming that in connected speech, the default form of tonal morphemes is full-tone syllables (called 'normal syllables'), and the crucial process in MC is not the stressing of certain syllables, but rather the de-stressing of 'normal' syllables – more appropriately described as prosodic weakening.

2.5 Intonation

Intonation is hereby defined narrowly as a subset of 'prosody', specifically referring to patterns of F0 changes. These patterns may be phonologically encoded in a tonal structure that is independent of lexical tones (Gussenhoven et al., 2013).

At this juncture, terminological clarification is needed regarding "tone". In autosegmental-metrical theory, pitch accents and *boundary tones* (such as H for high and L for low) are commonly used to represent pitch variations both within an utterance and at its final position (Liberman, 1975; Pierrehumbert, 1980; see also § 1.1.3). In this work, 'tone' refers instead to MC word-level *lexical tones* unless otherwise specified, while 'intonation' pertains to F0 variations across the entire utterance.

2.5.1 Global intonation trends

Prosodic downtrend phenomena have been observed to potentially occur physiologically, without conscious control by speakers, one of which is declination.

First introduced by Pike (1945), declination describes a general downward pitch trend throughout an utterance. The physiological nature of this phonetic effect is primarily attributed to a gradual decrease in subglottal pressure as air is expelled during speech. The extensive body of research on this topic, encompassing phonological descriptions of typologically diverse languages and grounded in statistical analysis, suggests that the phenomenon may be universal (Pierrehumbert, 1980; *inter alia*).

In tone languages, declination can occur also in mixed tone sequences; however, as one might expect, its existence and degree are most evident in phrases where all tones share the

same phonological value (Connell, 2001). Wang (2003) confirmed the presence of the declination effect in MC declarative sentences without new topics and focus, noting that downtrend effects become more noticeable in sentences with L tones or focus.

Fig. 16 illustrates the declination effect in MC. As reported, each PPh exhibits a global declining pitch trajectory superimposed upon the local pitch movements of lexical tones.

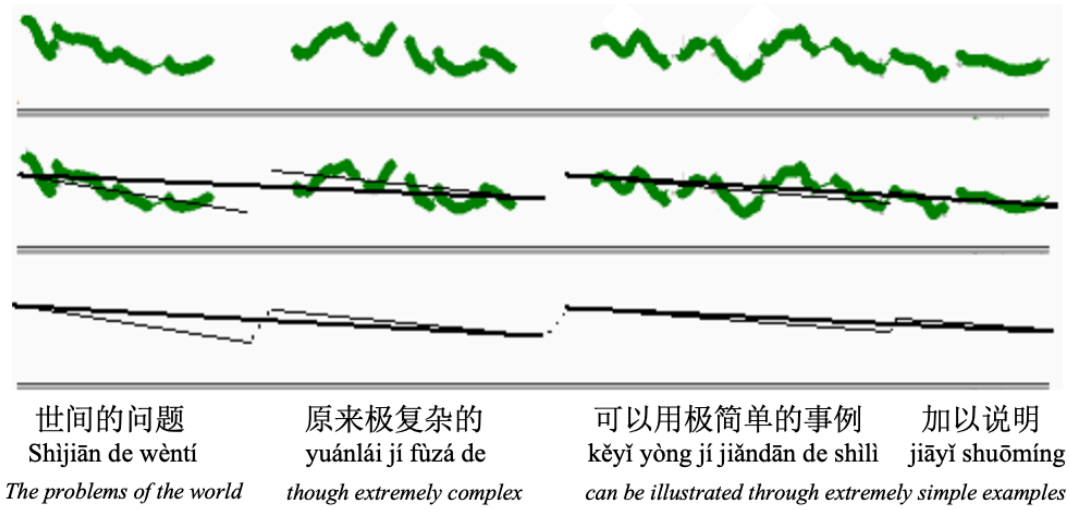


Figure 16 Declination effect in MC (from Cao 曹剑芬, 2016: 152)

In tone languages like MC, downtrend effects also include downstep phenomena, that has been described as the lowering of a H tone under certain specific conditions (Stewart, 1965). These effects may be classified as assimilatory carryover effects not strictly contingent upon inertia and articulatory speed constraints (Xu, Lee, 2022).

Within tonal phonology, downstep phenomena are conventionally divided into downstep proper (or non-automatic downstep) and downdrift, the latter generally interpreted in accordance with Stewart's (1965) notion of automatic downstep.

Downstep proper is evident when two H tone syllables occur consecutively, resulting in the lowering of the second H tone. Consequently, the sequence H H transforms into H$^{\downarrow}$H (Welmers, 1974; Connell, 2001). Studies of African tone systems indicate that, in most cases – if not universally – downstep proper results from the presence of an underlying ('floating') L tone, or from an L tone that has been lost historically (Stewart, 1965). Accordingly, the process may be represented abstractly as H (L) H.

The local assimilation between H and L tones is instead more evident in downdrift effect, described as the progressive lowering of the H tone following a L tone (see Connell, 2001).

Downdrift is exemplified in tone sequences such as H L (L) H, where the pitch of the second H tone is lower than that of the first H tone due to the influence of the intervening L tone syllable(s).

Fig. 17 provides an illustration of downdrift of an assimilatory type, resulting from the alternation of H and L tones, as reported by Connell (2001) based on Lindau's (1986) analysis of Hausa. Notably, the pitch slope in Fig. 17 is steeper than that observed under declination alone – approximately 33% per second – because the figure captures both the overall declination trend identified in the study and the local effect of downdrift, whereby H tones are progressively lowered in relation to adjacent L tones.

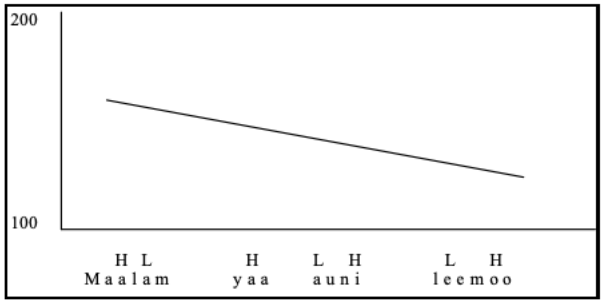


Figure 17 Illustration of downdrift in Hausa, showing alternating High (H) and Low (L) tones in the phrase "Maalam yaa auni leemoo" ('The teacher weighed the oranges'). From Connell (2001), adapted from Lindau (1986).

2.5.2 Tone and intonation interplay in MC

In Chinese languages, as in many others, F0 changes are the primary acoustic correlate of intonation (Chen, 2022, p. 345). Likewise, although lexical tones in MC exhibit intrinsic duration and intensity, the primary indicator for tonal contrasts remains the F0 contour (Whalen, Xu, 1992). Lexical tones in MC are used to differentiate word meanings, whereas intonation in both tone and non-tone languages is employed to convey syntactic and discourse-related information (Lin, 2007, p. 227); in addition, F0 is also exploited locally to mark focus. Hence, the concurrent use of the F0 channel for conveying both intonation and tone information presents the intriguing question of how utterance-level intonation and word-level lexical tones interact in MC.

Early discussions of Chinese intonation include mainly works on MC and Beijing Mandarin (Chao, 1968; Ho, 1977; Wang 王萍, Wu 吴宗济, 1982; Hu 胡明扬, 1987; Cao, 2002; Lin 林茂灿, 2006; Shi 石锋, 2011; *inter alia*) but also on other Chinese languages and dialects (Chang, 1958; Vance, 1976).

To date, both quantitative and qualitative data on Chinese intonation, particularly non-standard varieties, remain limited. However, the importance of research in this area has been considerably recognized, leading to an increasing number of descriptions of Chinese intonation and various attempts at intonation modeling. Several models have been proposed in the literature concerning how tone and intonation interact in MC (Chao, 1968; Wu 吴宗济, 1982; Gårding et al., 1983; Cao, 2002; Liu & Xu, 2005).

Chao's (1968) seminal account of tone-intonation interaction, conceptualizing lexical tones as small ripples superimposed on the larger waves of intonation, and his elastic-sheet analogy for pitch range expansion and compression in stress marking have exerted a lasting influence on subsequent theoretical models of Chinese intonation.

While Chao initially proposed that pitch register could be raised or lowered in a manner analogous to – yet independent from – the lengthening or shortening of syllable duration, Gårding et al. (1983) were the first to formalize pitch register manipulation in MC with the concept of the Range Grid. The authors argue that intonation in MC may be exemplified by a two-line grid: lexical tones are defined by pitch contours within specified upper and lower limits of the grid, whereas sentence intonation is marked by overarching rising and falling patterns, with adjustments occurring within the grid's range.

Shen 沈炯 (1992) expanded on the 'grid' concept by advocating for separate F0 lines: a gradually falling top line and a slightly undulating base line that ends at a much higher point to encode interrogativity. In fact, identification of tone tends to be relatively independent of intonation in MC, while the recognition of intonation seems to depend on the specific lexical tone. However, as noted by Chen (2022), this latter claim still requires further empirical investigation to be fully substantiated.

Such interactions between lexical tone and intonation observed in behavioral studies are supported by ERP data as well: Ren et al. (2009, 2013) demonstrated significant mismatch negativity (MMN) – an indicator of preattentive detection of acoustic changes – for T4 (falling contour) but not for T2 (rising contour) when distinguishing between questions and statements. Similarly, Liu et al. (2016) observed P300 waves – which reflect processes involved in stimulus evaluation or categorization – specifically when processing questions that conclude with T4, but not for questions ending with T2. In summary, the T2/T4 asymmetry reveals higher accuracy and quicker response times for T4 in identifying both declarative and interrogative intonation, consistent with Yuan's (2011) findings. This suggests that tone identification

contributes to the mapping of F0 contours onto intonational categories, underscoring an interaction between tone and intonation at the phonological level.

Unlike full tones, in MC the neutral tone appears more significantly influenced by intonation. Generally, the neutral tone merges the pitch value at the end of the preceding full tone with the pitch value dictated by intonation (Shen, 1990b). For instance, if the neutral tone's pitch is high after T3 (see post-low bouncing effect in § 2.3.3), it becomes even higher if the intonation at that point rises, such as at the end of a question.

According to Shen (1990a), 1) the question particle *ma* 吗, unlike other neutral-toned particles and syllables, always ends with a high pitch, regardless of the preceding syllable's tone (see also Casentini & Francolino, 2025 for a preliminary prosody-pragmatics comparative study on two *ma* question particles, i.e. ma1 吗 and ma2 嘛); additionally, 2) when multiple T0 syllables occur in sequence, the pitch contour of those following the initial occurrence tends to exhibit greater intonational flexibility. Subsequent T0 syllables are less constrained by the tonal specifications of preceding syllables and demonstrated increased susceptibility to intonational modulation (Fang & Xu, 2007; Lin, 2007).

2.6 Linguistic and paralinguistic functions of prosody: an overview of MC

In MC, it has traditionally been hypothesized that the use of sentence-final particles (SFPs), which encode specific syntactic and contextual meanings typically conveyed through intonation in non-tonal languages, reduces the need for significant pitch contour variation related to intonational patterns. However, from the functional viewpoint, intonation in Chinese appears as prevalent as in non-tonal languages.

Indeed, prosody in MC serves a wide range of linguistic and paralinguistic functions beyond the expression of lexical meaning. It encodes interrogativity (§ 2.6.1) and focus (§ 2.6.2), marks the prosodic structure of an utterance (§ 2.6.3 and § 2.6.4), signals discourse structure and regulates turn-taking (§ 2.6.5), communicate intentions (§ 2.6.6), expresses emotional states and attitudes (such as politeness and sarcasm) (§ 2.6.7), etc. (Chen, 2022, p. 337). Moreover, global intonation trends have been found in MC as well, as described earlier in this chapter (see § 2.5.1).

2.6.1 Prosodic encoding of questions in MC: a comparative perspective

Traditional universal models of interrogative encoding posit that, across languages, questions are typically characterized by an overall raised pitch level and/or a final rising pitch

movement. Some scholars have linked this tendency to the biological foundations of human speech, suggesting that rising pitch conveys meanings such as uncertainty or incompleteness (Ohala, 1983; Gussenhoven, 2016). However, the claim that such prosodic patterns constitute a universal mapping between form and meaning has been challenged (Rialland, 2007; *inter alia*), particularly in light of methodological differences across studies, including the use of read versus spontaneous speech. Notably, some studies highlighted that spontaneous yes-no questions frequently terminate with a L boundary tone, in contrast to the H or rising patterns typically observed in read questions within the same language varieties.

Specifically, rising-falling contour has been documented for yes-no questions in several Central Italian varieties, notably in Roman read speech. While in spontaneous speech, Giordano (2006) identifies a similar tendency, Sardelli and Marotta (2007) and Marotta and Sardelli (2009) propose an alternative interpretation, analyzing spontaneous Roman questions as exhibiting a falling-rising (H*+L H%) contour instead.

By contrast, in the Sienese variety, the falling-rising pattern (H+L* L-H%) is found to be predominant in read speech, whereas in spontaneous productions, approximately half of the questions display a rising-falling contour (Marotta & Sorianello, 1999).

The observed differences across varieties and speech styles indicate that intonational marking of interrogativity in Central Italian varieties is strongly conditioned by communicative context: read speech adheres to more conventionalized patterns, while spontaneous speech displays considerable prosodic variation and flexibility (see Savino, 2012 and references therein).

Research into how interrogativity is prosodically encoded in Chinese languages has predominantly concentrated on MC and Cantonese. In Cantonese, it has been widely demonstrated that the final F0 rise in questions has likely developed into a grammaticalized local H boundary tone (Lee et al., 2024), which serves as a distinct marker of question intonation but may obscure the original lexical tone contour (Chen, 2022, p. 349 and references therein).

On the other hand, questions in MC are identified by a globally raised F0, preserving lexical tonal contours (Ho, 1977; Shen, 1990a; Lin, 2007), as initially observed in the groundbreaking research by Chao 趙元任 (1933). Unlike Cantonese, MC does not appear to grammaticalize a final F0 rise as a local H boundary tone, since declarative and interrogative intonation patterns are generally differentiated through alternative prosodic cues (Yuan et al., 2002). With regard to MC, Shen (1990a) identified three fundamental intonation patterns (see Fig. 18, adapted and

simplified from Lin, 2007, p. 229): one associated with declarative statements and two with interrogatives, distinguished by a high or low utterance-final pitch.

Type 1, typically associated with declarative sentences (e.g., 他明天来 *tā míngtiān lái* 'he will come tomorrow'), begins at a mid pitch level, rises to mid-high, and concludes with a fall to low. Type 2, corresponding to interrogatives with a high utterance-final pitch (e.g., 他明天来？ *tā míngtiān lái?* 'he will come tomorrow?'), starts at mid-high, rises to high, undergoes a slight drop, and ends at high or mid-high. This type includes unmarked yes–no questions and particle questions; echo questions, although not explicitly discussed by the author, can also be accommodated within this category. Type 3, used for interrogatives with a low utterance-final pitch (e.g., 谁明天来？ *shéi míngtiān lái?* 'who will come tomorrow?'), begins at mid-high, rises to high, and then gradually falls to low. This category includes wh-questions, alternative questions, and A-not-A questions.

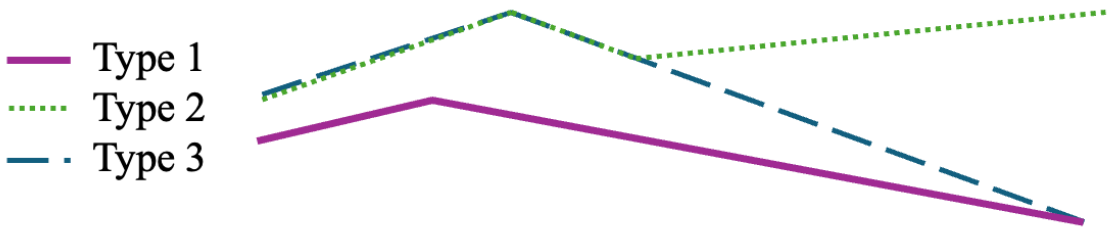


Figure 18 Basic types of intonation pattern according to Shen (1990a)

Lee (2005, p. 35) suggests considering both global and local F0 modifications in yes-no questions under different syntactic (e.g., with or without SFPs) and pragmatic (e.g., information-seeking question vs. echo question) conditions. Several studies suggest that ma-particle questions are prosodically distinct not only from statements but also from syntactically unmarked yes-no questions (Yang, 1995; Lee, 2005). These studies indicate that intonational manipulations are more pronounced in syntactically unmarked yes-no questions. For instance, Yang (1995) finds that although the global pitch register is raised in ma-particle questions, it is not as elevated as in syntactically unmarked yes-no questions.

Moreover, intonational patterns in certain syntactically marked question types seem to interact with focus structures. Narrow focus, typically accompanied by an expanded pitch range, is often realized on the question-word in question-word questions (Yi, 2004; Lee, 2005). This observation suggests that the traditional binary classification of MC intonation into declaratives and interrogatives is overly simplistic. A more comprehensive account must consider the

interplay between focus and question intonation – an area that remains relatively underexplored and is a primary concern of the present work.

Lin (2007, p. 231) argues that in emotive and expressive utterances, especially when emphasizing the final word, an additional tone feature may be added, as described by Chao 趙元任 (1933). In fact, despite the role of F0 raising over non-final syllables in encoding question intonation, some researchers argue that question intonation is primarily marked by a significant F0 rise over the final syllable (Chang, 1958; Hu 胡明扬, 1987). For this reason, some scholars have explicitly supported the necessity to represent question intonation in MC with a boundary tone (Peng et al., 2005; Lin 林茂灿, 2006), which either raises or lowers the lexical tones, such as [+RAISETONE] or [+LOWERTONE] as described by Lin 林茂灿 (2004). Jiang & Chen (2011) also embraced the concept of boundary tones, proposing two H boundary tones at the beginning and end of an interrogative utterance to account for both global and local F0 effects.

Liu & Xu (2005) confirmed the local and global prosodic marking of interrogativity in MC, alongside the dependence of various pragmatic question types on F0 modulation for interrogativity. In their study, listeners successfully distinguished between focus and yes-no questions (whether information-seeking or echo), particularly when the focus was not at the sentence-final position. They proposed that intonation components, such as focus and interrogativity, operate independently yet are encoded simultaneously within the global intonation framework. This theoretical approach is known as the Parallel Encoding and Target Approximation (PENTA) model of intonation (see also Xu, 2005), which posits that communicative meanings correlate with specific encoding parameters, i.e. local pitch targets, pitch range, articulatory strength, and duration. However, according to Chen (2022, p. 352), while acknowledging that tone and intonation typically function independently, PENTA doesn't seem to account for their interplay.

2.6.2 Prosodic encoding of focus in MC

The manipulation of pitch range of the focused item has been a central aspect across different models concerning the prosodic encoding of focus in MC and has been documented in numerous studies across various types of focus (Shih, 1988; Xu, 1999; Chen, Braun, 2006; Cao 曹文, 2010). Xu (1999)'s pivotal work significantly contributed to the hypothesis that focus is directly encoded through a tri-zone pitch range control mechanism: 1) minimal to no alteration before focus; 2) compression after focus, i.e. post-focus compression, or PFC; 3) in-focus expansion.

PFC phenomena have been observed both within and across other languages and dialects spoken in China, including Tibetan, Uygur, and Nanchang (Wang et al., 2011), as well as Lan-Yin Mandarin (Shen, Xu, 2016). Moreover, the PFC effect has been observed in several Indo-European and Altaic languages. There have also been observed dialects and languages across China where the PFC effect is absent. For instance, Cantonese (Wu, Xu, 2010), Taiwanese and Taiwan Mandarin (Xu et al., 2012), Wa, Deang, and Yi languages (Wang et al., 2011). Xu et al. (2012) proposed then a hypothesis suggesting that PFC may have a common origin, leading to a typological distinction between languages according to the presence of PFC. However, broader typological investigations on a larger scale are in order to support this claim.

As far as in-focus expansion is concerned, there are instances where different tones exhibit varying degrees of F0 range expansion to maintain their distinctiveness. For instance, it is plausible to draw parallels between Taiwanese (Pan, 2007) and Shanghai Chinese (Chen, 2009), where F0 modifications related to focus are robust only when they don't compromise the contrastive function of lexical tones. Such observations underscore the critical role of lexical tone characteristics in determining the extent to which focus-induced F0 range adjustments occur (Chen, 2022, p. 346).

With reference to the minimal domain of in-focus F0 modulation, evidences from both Wenzhou Chinese (Scholz, Chen, 2014) and Shanghai Chinese (Ling, Liang, 2017) indicates that the tone sandhi domain, rather than the syllable alone, functions as the minimal unit for conveying focus.

Although focus-induced F0 variation is generally distinct from tone-related F0 variation, their interaction is evident in dialects where F0 adjustments may be lacking or less pronounced (Chen, 2022, p. 346). Research on the perception of focus-induced F0 variation is relatively limited. Lee et al. (2016) demonstrated that speakers of MC can correctly identify focus with high accuracy (>90%) except when the focus is on a low-tone syllable (77%), possibly due to the limited F0 range manipulation potential of T3 under focus, particularly without additional cues like creakiness, as observed by Chen & Gussenhoven (2008) and Cao 曹文 (2010) .

Regarding the identification of lexical tones in MC under various focus conditions, Chen (2022, p. 347) initially observed an overall correct identification rate above chance levels. Notably, tones under focus were more accurately identified, regardless of their tonal context, i.e., whether presented with preceding and following tones or isolated with the target tone removed. The congruence of the preceding lexical tone, however, played a significant role. In congruent contexts (referred to as compatible contexts in Xu, 1994; see also the footnote in §

2.3.3 on the distinction between compatible and conflicting tone contexts), tones were more easily distinguished. By contrast, in conflicting contexts, the acoustic cues of the target tone were compromised by the carryover effect of the preceding tone, resulting in perceptual confusion – consistent with the findings reported by Xu (1994; see also § 2.3.3 for a review on carryover effects in MC).

Compared to lexical tones under focus, post-focus lexical tones exhibit higher sensitivity to both tonal context and congruency conditions. Consequently, Chen (2022, p. 348) reports that when post-focus tones were extracted from conflicting contexts and presented in isolation, there was a noticeable decrease in correct identification rates. Indeed, while lexical tones under focus typically display "hyperarticulation", post-focus lexical tones due to PFC effect are less prominent, often realized with a compressed F0 range and reduced articulatory effort, i.e., "hypoarticulated" (Chen, Gussenhoven, 2008; Chen, 2010, 2022).

Evidence from studies on MC suggests that contrastive focus relies chiefly on continuous pitch scaling mechanisms – such as F0 range expansion and register adjustment – rather than on categorical pitch accent assignment (Xu, 1999; Liu & Xu, 2005; Chen, 2006; Greif, 2010), a strategy more characteristic of non-tonal languages like Italian. Recent investigations reveal that focus marking in MC involves systematic pitch range adjustments, with focused constituents displaying expanded F0 excursions while maintaining lexical tone contours (Xu, 1999; Chen et al., 2009; Wang et al., 2020; Yang & Chen, 2021).

Cross-linguistic studies of prosodic transfer demonstrate asymmetrical acquisition patterns between Italian and other languages, particularly regarding accent placement and focus marking strategies (Bocci & Avesani, 2008; Avesani et al., 2015). Italian speakers learning L2 languages frequently exhibit negative prosodic transfer, characterized by inappropriate accent distribution that fails to align with target language pragmatic requirements.

L2 acquisition of tonal contrasts involves complex interactions between lexical tone requirements and phrase-level prosodic demands (Li, 2003; Chen, 2009; Wang et al., 2013; Yang, 2016). It can be hypothesized, therefore, that the maintenance of lexical tone integrity during focus marking may represent a particular area of difficulty for L2 learners from non-tonal language backgrounds, as it potentially involves competing demands between lexical and prosodic functions.

2.6.3 Prosodic phrasing

One key role of intonation is to delineate the prosodic structure of utterances, enabling listeners to efficiently parse and interpret the acoustic signal. Prosodic phrasing generally

reflects the syntactic structure of an utterance, although the correlation is not always exact (Shattuck-Hufnagel, Turk, 1996; Féry, 2016). This function is particularly evident when a sequence of words has multiple possible grammatical structures and requires prosody for disambiguation, as illustrated in examples (1) and (2) (Yang, 2016; Chen, 2022).

(1)	管	好	酒家
	guǎn	hǎo	jiǔjiā
	manage	well/good	restaurant

To manage a restaurant well / to manage a good restaurant

(2)	想念	水手	的	母亲
	xiǎngniàn	shuǐshǒu	de	mǔqīn
	miss	sailor	DE	mother

The mother who misses the sailor vs. to miss the sailor's mother

In speech, acoustic cues such as final lengthening, pause duration, and F0 variation help listeners disambiguate such structures (Tseng et al., 2005).

In example (1), grouping 'guǎn 管' and 'hǎo 好' together means "to manage a restaurant well," while grouping 'hǎo 好' and 'jiǔjiā 酒家' together means "to manage a good restaurant". Notably, 'guǎn', 'hǎo', and 'jiǔ(jiā)' all have an underlying low tone. In the first case, 'guǎn' undergoes tone sandhi and is realized with a rising F0 contour, whereas in the second case, 'hǎo' undergoes tone sandhi and surfaces with a rising F0 contour (see § 2.3.1 for the T3 sandhi rule). Thus, prosodic phrasing in this instance is indicated by prominent F0 information, as discussed by Chen (2022, p. 337).

Similarly, in (2), prosody can differentiate between the relative clause interpretation ("the mother who misses the sailor") and the verb phrase interpretation ("to miss the sailor's mother"). Indeed, behavioral and neurophysiological evidence further proved that intonation in MC plays a key role in helping listeners parse ambiguous utterances (Speer et al., 1989; Li, Yang, 2009; Li et al., 2011).

2.6.4 Information structure

Intonation serves a nuanced yet significant role in encoding the information structure of an utterance within its discourse context. In spoken interactions, speakers strive to convey

information effectively across discourse, a practice that varies across languages worldwide (see Féry, Ishihara, 2016). In several Chinese dialects, the concepts of focus and topic play crucial roles in information structure and are often reflected through prosody (see Chen, 2022, p. 338 and references therein).

Focus denotes the specific information that speakers highlight within discourse, typically new but not exclusively so (Krifka, 2008). This is exemplified in dialogues (3) and (4), where the same phrase *Mǎlì jiāo yǔyánxué* 玛丽教语言学 'Mary teaches linguistics' is uttered with distinct intonation patterns in response to queries about *what* Mary teaches (3a) versus *who* teaches linguistics (4a).

In (3b), {语言学}$_{\text{FOC}}$ is emphasized, indicating *linguistics* as the specific subject Mary teaches among alternatives. Conversely, in (4b), {玛丽}$_{\text{FOC}}$ is emphasized, signaling *Mary* as the unique individual teaching linguistics, distinguishing her from others (Chen, 2022, p. 338).

(3) a. 玛丽教什么？
Mǎlì jiāo shénme?
What does Mary teach?

b. 玛丽教{语言学}$_{\text{FOC}}$
Mǎlì jiāo yǔyánxué
Mary teaches {linguistics}$_{FOC}$

(4) a. 谁教语言学？
Shéi jiāo yǔyánxué?
Who teaches linguistics?

b. {玛丽}$_{\text{FOC}}$ 教语言学
Mǎlì jiāo yǔyánxué
{Mary}$_{FOC}$ teaches linguistics

Focal prominence in MC is conveyed through a combination of acoustic cues, including F0, duration (Chen, 2006; Chen & Gussenhoven, 2008), higher mean intensity (Chen et al., 2015) and hyper-articulated segmental contrast[10].

Analysis on the prosodic marking of different types of information structure demonstrated distinctions between focused new information and given information, as well as between informational focus and corrective focus (Chen, Braun, 2006; Wang, Xu, 2011; Ouyang, Kaiser, 2013). These prosodic differences reflect variations in cognitive processing patterns, as evidenced by experiments on eye-tracking (Chen et al., 2012) and brain responses (Chen et al., 2014). Focal prominence can alter the semantic interpretation of an utterance, as exemplified in (5) and (6) (Chen, 2022, p.339). In (5), emphasizing different parts of the utterance (such as 'John'约翰 versus 'apple' 苹果) can result in distinct truth-conditional meanings. For instance, if Mary gives both John and Peter an apple, the statement in (5) remains true when 'apple' 苹果 is emphasized but would be judged infelicitous if 'John' 约翰 is emphasized.

In (6), prosodic prominence clarifies ambiguities in semantic scope interpretation. Depending on where prominence is placed within the utterance, the aspect of what is 'surprising' can vary significantly. For example, emphasizing 'surprising' (*qíguài* 奇怪) suggests that the entire event of Ann getting married yesterday is unexpected. Conversely, placing prominence on 'yesterday' (*zuótiān* 昨天) would imply that it is the timing of the event that is surprising, etc.

(5)	玛丽	只	给	了	约翰	一	只	苹果
	Mǎlì	zhǐ	gěi	le	Yuēhàn	yī	zhī	píngguǒ
	Mary	only	give	ASP	John	one	CL	apple

Mary only gave John an apple

(6)	很	奇怪	安妮	昨天	结婚	了
	Hěn	qíguài	Ānnī	zuótiān	jiéhūn	le
	Very	surprising	Ann	yesterday	marry	ASP

It is surprising that Ann got married yesterday

[10] A more detailed discussion on how F0 modifications encode focus across some Chinese dialects can be found in Section 16.3 in Chen (2022).

On the other hand, topics are entities that, belonging to the common knowledge of the listeners, frame the information shared by a speaker. In example (7b), 'Mary' 玛丽 and 'Ann' 安妮 are identified as contrastive topics, meaning they are set in contrast to each other.

(7) *a.* 我知道玛丽和安妮都是老师。他们是教什么的？
Wǒ zhīdào Mǎlì hé Ānnī dōu shì lǎoshī. Tāmen shì jiāo shénme de?
I know that both Mary and Ann are teachers. What do they teach?

b. [玛丽]$_{TOP}$ 教语言学，[安妮]$_{TOP}$ 教地理
Mǎlì jiāo yǔyánxué, Ānnī jiāo dìlǐ
[Mary] $_{TOP}$ teaches linguistics, and [Ann]$_{TOP}$ teaches geography

Contrastive topics are realized in speech through prosodic prominence, similarly to how focal information is encoded. However, research suggests that contrastive topics may exhibit a distinct pattern of pitch contour: they typically display a noticeable rise in pitch followed by a gradual decline throughout the utterance (Chen, 2022, p. 340). This pattern seems to differ from the more immediate drop-off observed in non-contrastive topics, as attested for Shanghai Chinese (Chen, 2009) and MC (Wang, Xu, 2011). However, further investigation is needed to fully understand and attest this trend.

MC is well attested as a topic-prominent language (Li, Thompson, 1989; Morbiato, 2020). Various syntactic structures have been identified as topic-marking constructions (Xu, 2006), some of which are illustrated in examples (8) to (10). In these cases, topics are marked by a prosodic break, which in written texts may be indicated by punctuation.

(8) [玛丽]$_{TOP}$ 她教语言学
Mǎlì tā jiāo yǔyánxué.
Mary, she teaches linguistics

(9) 我[这个包]$_{TOP}$ 给我妹买的
Wǒ zhège bāo gěi wǒ mèi mǎi de
I bought this bag for my sister

(10) [水果]$_{TOP}$， 我最喜欢西瓜

Shuǐguǒ, wǒ zuì xǐhuān xīguā

(As for) fruits, I like watermelon the most

2.6.5 Discourse structure and turn-taking

To date, there has been relatively limited research examining the role of intonation in delineating discourse boundaries and signaling dialogue turns (Chen, 2022, p. 341). Tseng et al. (2005) identified PG in monologues, marked by duration, intensity, and F0 cues.

At the start of a PG, F0 resets and then gradually declines, with a significant drop at the end. Yang & Yang (2012) provided a detailed analysis of prosody in monologues, suggesting that various acoustic cues (including duration, F0 max/min, and pitch range) reliably indicate the rhetorical structure of the speech.

Levow (2004) demonstrated that in complex stories, topic endings correlate with final lengthening and a significant drop in pitch and intensity. Conversely, experimental studies on shorter discourses yielded evidence that a raised max F0 reliably indicates the start of a new topic (Wang, Xu, 2011; Yang, Yang, 2012).

In dialogues, intonation is crucial for managing turn-taking: for instance, Levow (2005) found that new turns typically begin with higher F0 and intensity, and speakers often elevate their initial F0 onset to interrupt and take over a turn.

2.6.6 Intention and interpretation

The view that speech production is an intentional action aimed at influencing the listener's behavior has a long tradition (see Grice, 1957; Austin, 1962; Searle, 1969). According to action theories of language, an utterance's meaning includes both the propositional content (locutionary force) and the speaker's intended effect (illocutionary force). Indeed, during speech, identical utterances may convey different illocutionary forces depending on the intonational patterns with which they are produced. For instance, a rising contour is often associated with polite requests, a falling contour with declarative statements, and a falling contour combined with an overall raised pitch with imperative force. It should be noted, however, that such associations are culturally mediated and therefore not universally applicable.

Additional prosodic and phonetic cues also contribute to the decoding of illocutionary force in MC utterances, including variations in intensity, speech rate, and segmental articulation (Hu

胡明扬, 1987; Shen 沈炯, 1992). In written language, such meanings can be partially conveyed through punctuation, as illustrated in examples 11-13. In more complex cases, however, the multifunctional nature of intonation makes such nuances difficult to represent in written language solely through punctuation, or even through paralinguistic markers such as emojis.

(11) 你把衣服穿上？
Nǐ bǎ yīfu chuān shàng?
Would you put on your clothes?

(12) 你把衣服穿上。
Nǐ bǎ yīfu chuān shàng.
Please put on your clothes.

(13) 你把衣服穿上！
Nǐ bǎ yīfu chuān shàng!
Put on your clothes!

Speakers often use prosody to clarify and make more explicit their intentions. For example, the phrase *duì bù duì* 对不对 ('right or not') can function both as a tag question or a pragmatic marker, depending on its prosodic realization. Chen & He (2001) found that the way the teacher employs prosodic cues within this phrase actually influence students' responses. Chen (2022) suggests that future research should explore how prosody conveys various speech acts, such as requests, complaints, warnings, promises, and apologies in MC.

The most studied aspect of speech act marking is using intonation to distinguish assertions (constative speech acts) from inquiries (classified among directive speech acts).

In most of the world's languages, declarative and interrogative sentences are distinguished through syntactic structure, word order, the presence of interrogative elements, or the use of particles. MC includes modal particles like ma 吗, ba 吧 and ne 呢 to signal questions, though these are not mandatory especially in colloquial speech. As a result, it happens that unmarked interrogative sentences syntactically resemble declaratives, not only in echo questions but also in unmarked yes-no questions (14) and in wh-questions (15).

(14) 王先生去北京了

Wáng xiānsheng qù Běijīng le

Did Mr. Wang go to Beijing? OR *Mr. Wang went to Beijing.*

(15) 王先生没去哪里

Wáng xiānsheng méi qù nǎlǐ

Where didn't Mr. Wang go? OR *Mr. Wang did not go anywhere.*

2.6.7 Emotion and attitude expression

Intonation also provides insight into interlocutors' emotions and attitudes, functioning as a paralinguistic aspect of prosody. Cross-linguistic studies indicate that core emotions are acoustically encoded in similar ways across languages, allowing non-native listeners to recognize them above chance level (Pell et al., 2009; Cheang, Pell, 2011; Liu, Pell, 2012; Paulmann, Uskul, 2014). This suggests that emotional states may have universal prosodic features due to their biological and cognitive underpinnings (Ohala, 1984; Gussenhoven, 2004; Xu et al., 2013). However, as linguistic systems vary, the paralinguistic expression of emotions may be shaped by the specific linguistic constraints of each system (Scherer et al., 2001).

Experimental studies on emotional intonation in MC are still limited but growing (Chen, 2022). Liu & Pell (2012) investigated how MC speakers produce and perceive seven emotions: anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality. They found that anger and surprise are associated with higher mean F0 and greater amplitude variation compared to happiness. In contrast, sadness, disgust, fear, and neutrality exhibit lower mean F0 and smaller amplitude variations. Speech rate and harmonics-to-noise ratio (HNR) [11] also vary systematically with emotion: anger and fear are characterized by a higher speech rate, while a slower speech rate distinguishes sadness and disgust; moreover, sadness has a high HNR compared to low HNR for anger and disgust. The acoustic measures demonstrated strong reliability in distinguishing among the seven emotions, indicating that different vocal emotions are associated with distinct acoustic profiles.

Li et al. (2011) studied how lexical tones are realized under seven emotional categories using data from two speakers. They found that sadness and disgust are expressed with a lower

[11] HNR is an acoustic measure in voice analysis that quantifies the balance between periodic (harmonic) and non-periodic (noise) components in a sound signal, expressed in decibels (dB). A high HNR indicates predominant harmonic energy and vocal clarity, while a low HNR signifies substantial noise content, often associated with breathiness, roughness, or voice pathology.

pitch register compared to neutral and fear, while happiness and surprise feature a more expanded pitch range and raised pitch register. Anger was realized differently across speakers: the female speaker employed a reduced pitch range, whereas the male speaker produced an expanded range. This divergence suggests that there is no straightforward correlation between pitch range or register and the expression of anger. Additionally, speakers added pitch targets to lexical tonal contours, with rising F0 for surprise and happiness, and falling F0 for disgust, anger, and fear. Chen (2022, p. 344) associates these emotion-induced pitch modifications to that observed by Chao 趙元任 (1933) and Mueller-Liu (2006).

Particularly relevant in this regard is the study by Zheng et al. (2025), which investigated whether the pitch characteristics of MC lexical tones exert an iconic influence on emotional experience. Drawing on three complementary approaches – two corpus-based analyses and an affect rating experiment involving both real and nonce disyllabic words – the authors examined the relationship between tonal sequences and affective responses. Their results demonstrated that falling-falling tonal patterns were consistently rated as more arousing, while high-high and rising-rising sequences were more frequently associated with positive valence. These findings indicate that even lexical tones, in addition to intonation, may exhibit subtle affective iconicity, thereby influencing the emotional perception of tone-carrying words. Emotions reflect fundamental internal psycho-physiological states of speakers, whereas attitudes convey higher-level social information. Commonly discussed attitudes include sarcasm, sincerity, humor, confidence, friendliness, and politeness (Wichmann, 2000; Pell, 2006).

Prosody plays a crucial role in marking attitudes across languages, although the specific acoustic parameters used to convey each attitude may differ. For instance, sarcasm in Cantonese appears to be correlated with a raised F0 but reduced amplitude and F0 range (Cheang, Pell, 2009), whereas in English, it is signaled by lowered F0. Sarcasm in Cantonese has also been associated with a slower speech rate. Humor, on the other hand, is marked by significantly lower HNR compared to other attitudes. Cheang & Pell (2011) extended this research, supporting the view that Cantonese speakers can accurately distinguish between different attitudes using prosodic cues. Regarding MC, Li & Wang (2004) demonstrated that friendliness is primarily expressed through a higher pitch. Gu et al. (2011) observed differences in both speech rate and F0 across various attitude pairs such as friendly-hostile, polite-rude, serious-joking, praising-blaming, and confident-uncertain. Lu et al. (2012) corroborated the role of prosody in attitude expression in MC; however, their analyses suggest that there is not necessarily a direct one-to-one relationship between specific prosodic patterns and attitudes.

3. Methods

Building on the theoretical frameworks and the bibliographic review outlined in Chapters 1 and 2, the present research project was designed to investigate Italian university learners' production of Mandarin lexical tones within minimal intonational units. The project consists of three interrelated empirical studies, each addressing how Italian learners of Mandarin encode higher-level prosodic information – specifically contrastive focus and sentence-type intonation (declarative vs. echo question) – within minimal prosodic domains, namely disyllabic phrases.

Specifically, Study 1 (see § 4) establishes a baseline by assessing tone identification and production in isolated monosyllabic and disyllabic target words. Study 2 (see § 5) investigates the prosodic encoding of contrastive focus in disyllabic statements embedded in short dialogues. Study 3 (see § 6) examines sentence-type intonation, focusing on the production of declaratives and echo questions. The stimuli for Studies 2 and 3 were selected as subcorpora from the main corpus (§ 3.3) and were segmentally matched to the disyllabic target words from Study 1, elicited separately. This design allowed us to control for potential segmental variability (see § 2.3.4).

A central issue addressed in these studies concerns whether – and how – L2 learners' prosodic modulation of tone within intonational phrases diverges from native-speaker patterns, and whether their tonal implementation remains primarily lexically governed and comparatively insensitive to discourse-level variation. In addition, to examine potential predictors of tone-intonation accuracy in L2 Mandarin, we considered not only learners' academic year, but also their Mandarin oral proficiency and musical aptitude, given the positive relationship between L2 prosodic ability and musicality reported in the literature (§ 1.1.4).

In fact, the overarching objective is to identify systematic patterns of prosodic misalignment that extend beyond the misproduction of individual lexical tones and that may reveal broader difficulties in mapping tonal targets onto intonational structures. This focus is informed by recent research demonstrating that preserving lexical tone accuracy while marking focus constitutes a major difficulty for L2 learners whose first language lacks lexical tone (see § 2.6.2), and that intonational patterns interact with focus structures in non-trivial ways (see § 2.6.1).

To this end, the experimental items were designed to simultaneously encode contrastive focus and sentence-type intonation. The decision to employ syntactically unmarked echo questions is motivated by two considerations. First, these question types are especially sensitive to intonational modulation, making them an ideal context in which to preliminary investigate

the interaction between lexical tones and sentence-level intonation. Second, they allow for the analysis of phrase-final full-tone syllables, which in MC questions are generally characterized by global F0 patterns while preserving tonal contours (see § 2.6.1). Moreover, syntactically unmarked questions are common in colloquial MC and are not limited to echo-question contexts (see § 2.6.6).

The dataset includes F0-based acoustic measures extracted from the productions of Italian L2 learners of Mandarin. A control group of native Mandarin speakers was also included, both to ensure comparability with the findings reviewed in Chapter 2 and to provide a benchmark for evaluating L2 productions.

As noted in Chapter 2, prior research indicates that tone identification in MC is less accurate when focus falls on low-tone syllables, particularly Tone 3 (T3), whose F0 range remains restricted under focus unless reinforced by secondary cues such as laryngealization or creakiness (Chen & Gussenhoven, 2008; Cao 曹文, 2010; Lee et al., 2016). Owing to these complexities, T3 was included in Studies 1 and 2 primarily as a control tone, but excluded from Study 3, which required simultaneous prosodic encoding of sentence type and focus, and was therefore only partially addressed in the general discussion.

The analyses rely on well-established acoustic parameters for MC tone and intonation (see § 2.5.2), confirmed for reliability in two pilot studies (Francolino, 2024a; Francolino & Cao, 2024). Generalized Additive Mixed Models (GAMMs) were employed for contour analyses, while Generalized Linear Mixed Models (GLMMs) were used for the analysis of other curve parameters. Estimated Marginal Means (EMMs) were subsequently computed for post-hoc group comparisons.

In the final Chapter (§ 7), the results from the three studies will be interpreted in light of the L2 Intonation Learning Theory (see § 1.1.3), in order to evaluate how Italian learners of Mandarin negotiate the interaction between lexical tones and intonational structures.

For the sake of transparency and reproducibility, a detailed HTML report containing the full statistical analysis is available upon request.

3.1 Participants

A total of 42 Italian learners of Mandarin Chinese (M = 21.7 years, SD = 1.46) were recruited from two institutions: the University for Foreigners of Siena and Roma Tre University.

Participants were enrolled across three academic levels – second-year BA, third-year BA, and first-year MA (see Tab. 1). The BA students were enrolled in the Linguistic Mediation

curriculum (L-12, *Mediazione linguistica*), while the MA students were enrolled in the Modern Languages and Literatures curriculum (LM-38, *Lingue e letterature moderne*) or the Linguistics curriculum (LM-39, *Linguistica*). All participants completed a multisectional post-test questionnaire (see § 3.4.1). All participants were native Italian speakers, with the majority born and raised either in the Rome or Siena area, and none reporting hearing or speech impairments. Each provided written informed consent (see Appendix A) and voluntarily participated in the study in exchange for a compensation of €8.

Table 1 Distribution of L2 Participants by University and Academic Level

University	BA2	BA3	MA1	Total
University for Foreigners of Siena	8	10	8	26
Roma Tre University	6	4	6	16
Total	14	14	14	**42**

An additional control group of 10 native Mandarin speakers (Mean age = 21.3 years; SD = 2.22) also participated in the study. All were university students and held a Putonghua Shuiping Ceshi (PSC) Level 1-B certification, indicating high proficiency in MC. The control group consisted of students enrolled at Beijing Language and Culture University, the majority of whom originated from northern China and had no proficiency in Italian. None reported hearing or speech impairments. All participants volunteered to take part in the study, completed a brief language background questionnaire (see § 3.4.2), and were compensated ¥100 for their participation.

Table 2 Background Information for Native Mandarin Speakers

Speaker	Age	Degree	Provenance
Ch1	20	English	Jilin
Ch2	20	Linguistics	Hebei
Ch3	20	Psychology	Tianjin
Ch4	19	Chinese	Anhui
Ch5	21	English	Anhui
Ch6	25	Literature	Jiangxi
Ch7	19	Chinese	Henan
Ch8	24	L2 Mandarin Teaching	Guangdong
Ch9	21	Computer Science & Tech.	Henan
Ch10	24	German Translation	Shandong

The inclusion of a native control group enabled systematic comparison of L1 and L2 prosodic patterns and provided a benchmark for evaluating the tonal accuracy and focus-marking strategies of the L2 speakers.

Reasonably, the control group was excluded from the pre-test procedures described in the following section.

3.2 Pre-test Procedures

Prior to the main reading task, participants underwent a preliminary oral competence assessment including two components: (i) tone identification and production tasks, and (ii) an oral competence assessment designed to evaluate general speaking skills. A brief instruction session preceded both components to ensure participants were familiar with the procedures and expectations.

3.2.1 Tone identification task

The tone identification task was administered online and required participants to identify the tonal value of syllables in monosyllabic and disyllabic Mandarin words. All stimuli were produced by a pre-recorded native speaker of MC. Each audio stimulus was played twice by default, after which participants were prompted to provide an oral tonal identification of the target syllable(s). If needed, participants could request one additional playback before responding.

The stimulus set included 32 randomized items, including 16 monosyllabic words (4 per tone) and 16 disyllabic words (1 per tone combination).

Target syllables featured either sonorant or voiceless initials, and nasal codas were excluded to avoid interference with tone perception. All words were mid-frequency Mandarin lexical items, selected from the *Xiàndài Hànyǔ Chángyòng Cíbiǎo* 现代汉语常用词表 (2021), with a mean frequency of 1836 (SD = 1581), and maximum frequency below 7074.

Two native Mandarin speakers also completed the task as a control group to establish baseline performance.

Table 3 Monosyllabic Stimuli by Tone and Segmental Class

	yao	***tu***	***ji***	***fei***
T1	yāo 腰 (‘waist’)	tū 突 (‘protrude’)	jī 鸡 (‘chicken’)	fēi 飞 (‘fly’)
T2	yáo 摇 (‘shake’)	tú 徒 (‘disciple’)	jí 急 (‘urgent’)	féi 肥 (‘fat’)
T3	yǎo 咬 (‘bite’)	tǔ 土 (‘soil’)	jǐ 几 (‘how many’)	fěi 匪 (‘bandit’)
T4	yào 药 (‘medicine’)	tù 兔 (‘rabbit’)	jì 计 (‘plan’)	fèi 费 (‘cost’)

Table 4 Disyllabic Stimuli by Tone Combination

	huā 花 (‘flower’)	**xié 鞋 (‘shoes’)**	**shǒu 手 (‘hand’)**	**piào 票 (‘ticket’)**
hēi 黑 (‘black’)	hēihuā 黑花	hēixié 黑鞋	hēishǒu 黑手	hēipiào 黑票
bái 白 (‘white’)	báihuā 白花	báixié 白鞋	báishǒu 白手	báipiào 白票
zǐ 紫 (‘purple’)	zǐhuā 紫花	zǐxié 紫鞋	zǐshǒu 紫手	zǐpiào 紫票
lǜ 绿 (‘green’)	lǜhuā 绿花	lǜxié 绿鞋	lǜshǒu 绿手	lǜpiào 绿票

3.2.2 Tone production task

The tone production task was administered in person, immediately preceding the main reading task, in a one-on-one session with the researcher. Participants were asked to produce a randomized list of 16 monosyllabic words (see Tab. 5) and 16 disyllabic Mandarin phrases (see Tab. 6), each presented in isolation. A brief pause and an auditory beep followed each stimulus, designed to reduce potential list effects and maintain a consistent elicitation rhythm.

The disyllabic stimuli comprised all target phrases from the main reading task (see § 3.3.1), thereby enabling a direct comparison between isolated and contextualized productions. Productions elicited in isolation were used as a control condition in the analyses. All items were selected from mid-frequency vocabulary, based on data from the *Xiàndài Hànyǔ Chángyòng Cíbiǎo* 现代汉语常用词表 (2021), with a mean frequency of 2344 (SD = 2278) and individual frequencies below 8777.

Table 5 Monosyllabic Stimuli by Phonological Class

	All Sonorant	Nasal Codas	Palatal Initials	Plosive Initials
T1	mā 妈 (‘mother’)	chūn 春 (‘spring’)	qī 七 (‘seven’)	bāo 包 (‘bag’)
T2	ná 拿 (‘take’)	cháng 常 (‘often’)	jí 极 (‘extreme’)	tiáo 条 (‘strip’)
T3	mǎ 马 (‘horse’)	jiǎng 讲 (‘speak’)	xiǎo 小 (‘small’)	bǎo 饱 (‘full’)
T4	wàn 万 (‘ten thousand’)	bàn 半 (‘half’)	xiào 笑 (‘laugh’)	dà 大 (‘big’)

Table 6 Disyllabic Stimuli Tone Combinations

	T1	T2	T3	T4
T1	hē zhōu 喝粥 (‘drink porridge’)	hē chá 喝茶 (‘drink tea’)	hē jiǔ 喝酒 (‘drink alcohol’)	jiā cài 加菜 (‘add dishes’)
T2	xué chē 学车 (‘learn to drive’)	dú bó 读博 (‘pursue PhD’)	tí shuǐ 提水 (‘carry water’)	zhái cài 择菜 (‘pick vegetables’)
T3	zhǔ zhōu 煮粥 (‘cook porridge’)	zhǔ chá 煮茶 (‘boil tea’)	xǐ jiǎo 洗脚 (‘wash feet’)	zhǔ cài 煮菜 (‘cook vegetables’)
T4	jiè shū 借书 (‘borrow books’)	dài chá 带茶 (‘bring tea’)	dài jiǔ 带酒 (‘bring alcohol’)	zuò cài 做菜 (‘cook dishes’)

3.2.3 Oral competences assessment

The general oral competence assessment comprised two online tasks, both adapted from the HSKK Intermediate proficiency exam. These tasks were designed to evaluate participants’ spontaneous speaking ability, fluency, and lexical range in Mandarin.

In the first task, participants were presented with two visual prompts and asked to select one. They were then required to deliver a spontaneous spoken response of approximately two minutes. Rather than providing a literal description of the image, participants were encouraged to offer personal reflections or comments, as long as they remained thematically relevant (e.g., travel, study, work, reading, etc.).

Participants were given one minute of preparation time before responding. During this time, they were allowed to organize their thoughts and, if needed, consult an online dictionary to support vocabulary selection.

Figure 19 Visual prompts used in Task 1 ("*Choose picture A or B and, after one minute of preparation, speak about it for approximately two minutes*").

The second task mirrored the structure of the first, with the exception that participants were given two written prompts instead of visual stimuli. After selecting one prompt and a one-minute preparation period, they delivered a spoken response of roughly two minutes in duration. The written prompts covered similar everyday topics and were designed to elicit structured, yet spontaneous speech that reflected participants' ability to reason, describe, and express opinions in Mandarin.

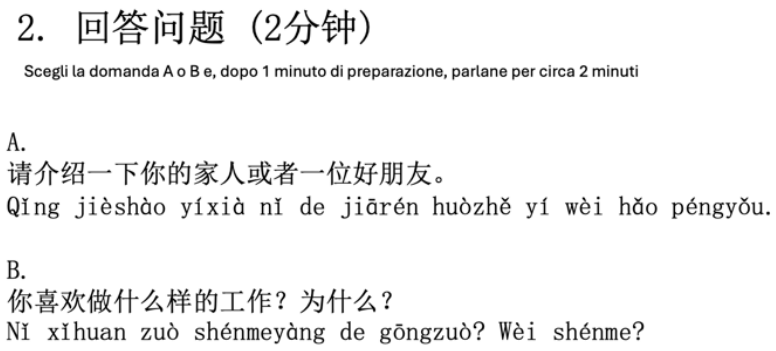


Figure 20 Written prompts in Task 2 ("*Choose question A or B and, after one minute of preparation, talk about it for approximately two minutes.")*

To assess participants' overall oral proficiency, two trained native speakers of Mandarin, both serving as instructors in Italian universities, evaluated the anonymized recordings. Rating criteria were adapted from the HSKK Scoring Guidelines and aligned with the CEFR (2018). Prior to data collection, the rubric was vetted by two international expert in Mandarin pronunciation assessment, and a calibration session was held to harmonize scoring practices. The final rubric included four core dimensions, provided in the Tab. 7 below.

Table 7 Oral proficiency rating criteria (Adapted from HSKK and CEFR guidelines and hereby translated from Mandarin)

Dimension	Elementary (A1-A2)	Intermediate (B1-B2)	Advanced (C1-C2)
Content richness & effectiveness	Fragmented information; lacks context or abstraction.	Relevant but concrete; limited abstract content.	Integrates concrete and abstract content with nuance.
Fluency	Frequent pauses and difficulty maintaining coherence.	Generally fluid, with occasional disruptions.	Smooth, spontaneous delivery with minimal hesitation.
Phonological accuracy	Frequent mispronunciations impede comprehension.	Mostly intelligible, though systematic errors occur.	Clear, native-like intelligibility; minimal accent.
Lexicon and grammar	Basic vocabulary and simple syntax dominate.	Limited complexity, with basic cohesive devices.	Advanced structures and rich vocabulary used effectively.

3.3 Main task

The primary experimental task was a semi-scripted reading activity carried out in pairs. To minimize variation related to academic exposure and to ensure a baseline level of interlocutor familiarity, each participant was paired with a peer from the same university year.

The task comprised 65 short dialogues in total, including both target and filler items. Each dialogue was read twice in a non-consecutive order, ensuring that participants alternated speaker roles across repetitions. Consequently, the total number of pseudo-randomized dialogue tokens per participant amounted to 130.

Target dialogues embedded the critical experimental items (i.e., disyllabic phrases representing all relevant tone combinations), while filler dialogues were syntactically varied to avoid pattern anticipation or tonal priming.

The reading session was divided into four blocks of approximately 30 dialogues each, with three-minute rest intervals between blocks to reduce participant fatigue. During the breaks, participants were instructed to hydrate in order to preserve vocal quality and minimize the occurrence of unintended creakiness in their speech.

3.3.1 Target phrases design

In Modern Chinese, the majority of words are disyllabic (Huang & Liao, 2017). Wu Zongji (1982, 2004) demonstrated that disyllabic phrases constitute the basic intonational units of MC, suggesting that even short disyllabic sequences can exhibit complete and complex prosodic realizations. Cao 曹文 (2010) further demonstrated that in bi- and trisyllabic tonal

combinations occurring in context, prosodic variation involves not only local tonal alternations (e.g., tone sandhi) but also global modulations of phrasal prominence, thereby determining the intonational contour of the utterance. Cao's study also revealed that tonal variations at the phrasal level display a degree of recursivity, which can be described through generalizable prosodic models, at least for short utterances.

For these reasons – along with the need to minimize global prosodic phenomena such as declination (see § 2.5.1) and to facilitate articulation for student participants – the present study employed disyllabic target phrases.

All target items were disyllabic Verb-Object (VO) phrases, embedded as stand-alone utterances within the dialogue. This syntactic control served two purposes, namely ensuring uniformity in prosodic phrasing across conditions; and allowing the tonal contour of each syllable to be isolated and measured consistently.

In addition, all target syllables began with voiceless initials, thereby limiting the subsequent F0 analysis to the vocalic nucleus. The design encompassed all possible disyllabic combinations of full lexical tones, thus excluding the neutral tone. As already mentioned, the target phrases were segmentally identical to the stimuli employed in the tone production task (see § 3.2.2).

Tab. 8 lists the 16 disyllabic tone combinations used in the experimental task, along with the lexical items selected and their respective word frequency values, based on the *Xiàndài Hànyǔ Chángyòng Cíbiǎo* (现代汉语常用词表, 2021):

Table 8 Disyllabic Target Phrases and Lexical Frequencies

Tone Comb.	Target Phrase	Syl1 Wordfreq	Syl2 Wordfreq
T1T1	喝粥 *hē zhōu*	2637	8111
T1T2	喝茶 *hē chá*	2637	2045
T1T3	喝酒 *hē jiǔ*	2637	1135
T1T4	加菜 *jiā cài*	543	1146
T2T1	学车 *xué chē*	239	1561
T2T2	读博 *dú bó*	651	8777
T2T3	提水 *tí shuǐ*	779	143
T2T4	择菜 *zhái cài*	6588	1146
T3T1	煮粥 *zhǔ zhōu*	5362	8111
T3T2	煮茶 *zhǔ chá*	5362	2045
T3T3	洗脚 *xǐ jiǎo*	1409	1067
T3T4	煮菜 *zhǔ cài*	5362	1146

Tone Comb.	Target Phrase	Syl1 Wordfreq	Syl2 Wordfreq
T4T1	借书 *jiè shū*	1000	228
T4T2	带茶 *dài chá*	178	2045
T4T3	带酒 *dài jiǔ*	178	1135
T4T4	做菜 *zuò cài*	77	1146

Each target phrase was embedded within a short dialogue designed to elicit four distinct prosodic contexts, manipulating both sentence type and focus position. Specifically, the design targeted two sentence modalities – declarative statements and unmarked echo questions – crossed with contrastive focus on either the first syllable (*Syl1*) or the second syllable (*Syl2*). This yielded the following four experimental conditions:

- Statement with contrastive focus on Syl1;
- Echo question with contrastive focus on Syl1;
- Statement with contrastive focus on Syl2;
- Echo question with contrastive focus on Syl2.

Each of the 16 disyllabic tone combinations appeared in all four conditions, yielding a total of 64 unique target phrases (16 tone combinations × 2 sentence types × 2 focus positions).

The excerpt below exhibit how two tokens of the target phrase (zuò cài 做菜, T4-T4) were embedded in a short dialogue to induce prosodic focus on the second syllable. The English gloss and underlined focus marking, not displayed to participants, are presented here for illustration only. A complete list of target dialogues, presented in non-randomized order, is provided in Appendix B.

Table 9 Example target dialogue embedding focus on Syllable 2 (T4T4)

Turn	Sentence Type	Focus position	Mandarin	Gloss
A	NA	NA	你要做菜还是做点心？ *nǐ yào zuò cài háishì zuò diǎnxīn?*	Do you want to make a dish or make a dessert?
B	Statement	Syl2	做菜。 *zuò cài.*	*Make a dish.*
A	Echo Q	Syl2	做菜？ *zuò cài?*	*Make a dish?*
B	NA	NA	对，做点心的话时间太长。 *duì, zuò diǎnxīn dehuà shíjiān tài cháng.*	Yes, making a dessert takes too long.

3.3.2 Recording equipment and environment

Speech data were collected with high-fidelity digital recording equipment in sound-attenuated environments, ensuring optimal signal quality for precise acoustic analysis.

For the Italian learner group, recordings were conducted using an Aston Origin condenser microphone, positioned at a distance of approximately 15-25 cm (≈6-10 inches) from the participant's mouth. The microphone was connected to a Focusrite Scarlett 2i2 (4th generation) audio interface. Audio was captured using Logic Pro X on a MacBookPro17,1, and all recordings were saved in .wav format with a sampling rate of 48 kHz and 24-bit resolution.

Recording sessions took place in a quite environment at both participating institutions. For both the individual (pre-test) and paired (main task) sessions, participants were seated at a distance of approximately 30 cm (≈12 inches) from a computer monitor. The monitor was remotely controlled by the researcher to present stimuli, minimizing participant distraction and ensuring experimental consistency. The full experimental procedure per participant lasted approximately 60 minutes.

For the native speaker control group, data were collected in the phonetic laboratory of Beijing Language and Culture University. In this setting, two Shure MVL lavalier microphones were used (one per speaker), positioned at the same 15-25 cm distance from the mouth. Participants sat approximately 50 cm (≈20 inches) from the monitor, which was likewise remotely operated by the researcher. Recordings were saved in .wav format at 48 kHz/24-bit. Each native speaker session lasted approximately 20 minutes.

3.3.3 Data extraction and annotation

Time-normalized F0 data for the pitch-contour analysis were extracted using the ProsodyPro script (Xu, 2013) in Praat (Boersma & Weenink, 2025). The script sampled ten equidistant points from the selected region of each syllable.

Syllable boundaries for sonorant segments were manually annotated using a combination of visual cues (waveform and spectrogram) and auditory inspection.

Specifically, onset boundaries were typically marked at the beginning of visible periodicity and formant structure, whereas offset boundaries were placed at the point of spectral reduction preceding consonantal closure or following the loss of voicing. Fig. 21 illustrates the Praat work window in which syllables were annotated.

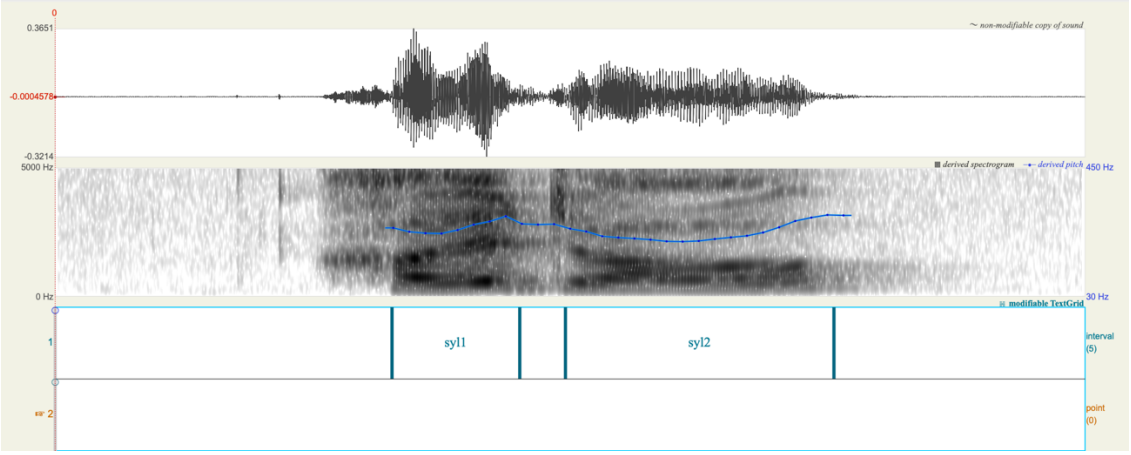


Figure 21 Praat work window showing syllable annotation based on the waveform and spectrogram

3.4 Post-test questionnaire

3.4.1 Experimental questionnaire design

The experimental questionnaire was administered in digital format and organized into four sections, each addressing a specific domain of individual variation.

These sections were designed to elicit information on participants' musical aptitude, previous language learning experience, and sociolinguistic background, thereby providing complementary variables for interpreting the experimental data through both quantitative and qualitative perspectives.

3.4.1.1 Section 1: Tone-Deafness Test

Participants began the questionnaire by completing the Tone-Deafness Test, an auditory pitch discrimination task hosted by the Music Lab at Yale University[12].

This citizen-science tool assesses pitch perception accuracy through a series of adaptive listening trials. Upon completion, participants were instructed to upload a screenshot of their test results before proceeding to the next section.

3.4.1.2 Section 2: Gold-MSI-IT (Musical Sophistication Index)

The second section consisted of the Italian version of the Goldsmiths Musical Sophistication Index (Gold-MSI-IT), a self-report inventory that assesses individuals' engagement with and

[12] https://www.themusiclab.org/quizzes/td (accessed Sep 20, 2025)

sensitivity to music (Müllensiefen et al., 2014; Santangelo et al., 2023). The questionnaire used 7-point Likert scales to measure musical behaviours and abilities across several domains:

- Affective engagement (e.g., frequency of emotional responses such as chills or nostalgia evoked by music);
- Technical competence (e.g., self-perceived accuracy in pitch and rhythm reproduction);
- Perceptual awareness (e.g., confidence in identifying vocal quality or musical errors);
- Musical activities (e.g., performance history, instrumental training);
- Exposure and listening habits (e.g., average hours of daily music listening, concert attendance).

In addition to qualitative measures, this section collected quantitative data on formal musical training (e.g., years of instruction, practice routines), enabling a richer profile of each participant's musical background.

3.4.1.3 Section 3: Mandarin Learning Background

The third section targeted participants' academic experience and language engagement with Mandarin (see Appendix C). Items captured:

- Institutional affiliation, academic year, and enrollment in university-level Chinese language courses;
- Self-assessed importance of key linguistic skills (grammar, writing, vocabulary, speaking) on a 5-point Likert scale;
- Attitudes toward pronunciation and focus on oral production in the classroom;
- Extracurricular learning behaviours, including private tutoring, self-study, and online learning;
- Language use frequency outside formal instruction (e.g., casual conversation, work contexts);
- Affective descriptors of the learning experience (e.g., "stimulating," "challenging," "frustrating").

This section was designed to assess both formal instruction and real-world usage, enabling a more holistic understanding of each learner's interaction with the target language.

3.4.1.4 Section 4: Sociolinguistic and Biographical Background

The final section collected background information relevant to individual variation (see Appendix D). Data points included:

- Demographic variables: age, residence, city mobility, duration of stay in specific countries or Italian regions;
- Language profile: L1(s), L2s, early exposure to Mandarin (e.g., high school coursework, immersion experiences), and overall language repertoire;
- Self-reported diagnoses of auditory impairments or learning difficulties (e.g., dyslexia), where applicable.

3.4.2 Control group questionnaire design

The native speaker control group completed a condensed sociolinguistic and academic questionnaire, focused on key identifiers such as Place of origin; Academic major; Foreign languages spoken (if any); and Speech/hearing status.

3.5 Learner Variables: Pre-Modelling Overview

This section presents the principal results connected to: 1) the operationalization of learner variables, specifically the Proficiency variable, derived from participants' performance on the HSKK speaking proficiency test and the Tone Identification Test; and the Musicality variable, based on scores from the Musical Sophistication Index and the Tone-Deafness Test; and 2) the pre-modeling screening of variables conducted to ensure their suitability for subsequent statistical analyses.

3.5.1 Learner variables construction

3.5.1.1 Proficiency variable

Participants' Mandarin proficiency was evaluated through two complementary approaches: (i) tone identification accuracy, and (ii) general oral production competence. These assessments provided both fine-grained phonological measures and global communicative indices, allowing for a multi-dimensional analysis of L2 Mandarin oral production ability.

Tone Identification Test

Before administering the test to learners, it was first validated by two native speakers, both of whom achieved a 100% accuracy rate. Students' results was first analyzed separately for monosyllabic and disyllabic target phrases. Subsequently, an overall score for the tone identification test was calculated by combining the results from both categories.

Monosyllabic target

First, an overall percentage score was calculated to assess accuracy for each tone. As outlined in the methods section (§ 3.2.1), each tone comprised four segments (*fei, tu, ji, yao*), which were presented to 42 participants, resulting in a total of 168 trials per tone. As reported in Tab. 10 below, T1 was identified with the highest accuracy, followed by T4, T2, and T3.

Table 10 Monosyllabic target identification test score per tone

Tone	Accuracy (%)
T1	91.7%
T4	83.9%
T2	57.7%
T3	54.2%

To assess specific misidentifications, a confusion matrix was computed. As reported in the heatmap below (Fig. 22), the most common confusions were between T2 and T3 pairs.

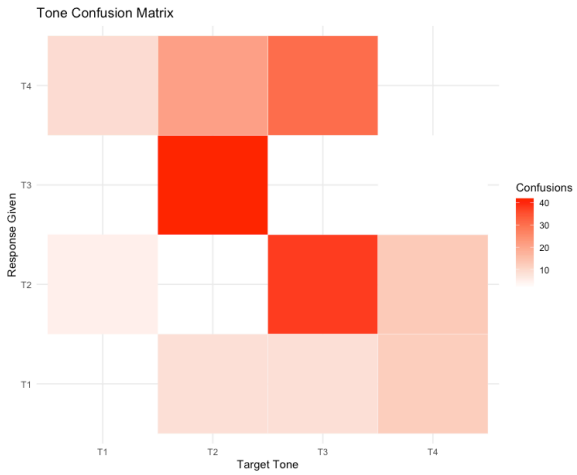


Figure 22 Monosyllabic target identification test confusion matrix

The average percentage of correctly identified tones was then computed for each student. Results reported substantial variation, ranging from 37.5% to 100%, reflecting heterogeneity in individual learner performance (see Appendix E).

A GLMM was fitted to model accuracy, incorporating Speaker and Item as random effects to account for inter-participant and inter-item variability. The results, summarized in Table 11, use T1 – the tone with the highest identification accuracy – as the reference level.

Table 11 Monosyllabic target identification test summary table

Term	Estimate	Accuracy %	Interpretation
(Intercept)	2.81	94.3%	Baseline log-odds of correct response for T1. Very high accuracy.
ConditionT2	-2.42	60%	Compared to T1, T2 is much harder to identify (significantly lower odds).
ConditionT3	-2.54	57%	T3 is significantly harder than T1 to identify
ConditionT4	-0.70	89%	T4 is slightly harder, but not significantly different from T1 (p = 0.32).

The model accounted for variation across both students and items, confirming that individual learner differences and item-specific characteristics both influenced performance (Speaker variance: 0.78; Item variance: 0.70). These findings support prior research that T2 and T3 are more confusable and harder for L2 Mandarin learners (Wang et al., 1999; *inter alia*). A likely explanation is that the high identification accuracy on T1 and relatively strong performance on T4 may derive from their acoustic salience and pitch contour distinctiveness. The confusion matrix and model-based estimates both corroborate the following hierarchy of difficulty in tone identification in isolated monosyllabic target words:

T1 ≈ T4 > T2 > T3

Disyllabic target

In the analysis of the disyllabic tone identification data, we explicitly controlled for the T3-T3 and T2-T3 sequences, both of which involve a first syllable that is phonetically realized as T2. Crucially, this syllable can be phonologically interpreted as either T2 or T3, depending on whether the learner considers the Mandarin third-tone sandhi (see § 2.3.1). To accommodate this dual possibility, we introduced an additional variable, termed OptCondition. This column specifies an alternative acceptable response in cases where ambiguity is expected (e.g., both T2 and T3 being valid targets). For items without such ambiguity, OptCondition was set to NA.

To evaluate accuracy in a manner aligned with this experimental design, we defined the binary variable Correct as 1 if a participant's response corresponded to either the primary Condition or, when applicable, the OptCondition; otherwise, it was coded as 0.

As an initial analysis, we computed each student's overall tone identification accuracy as the proportion of correct responses across all items (see Appendix E). We then calculated accuracy rate per Tone, and fit the results in the following table:

Table 12 Disyllabic target identification test score per tone

Tone	Accuracy (%)
T4	71.7%
T1	68.8%
T2	47.6%
T3	32.7%

Results reveal that T3 yielded the lowest recognition accuracy, indicating greater perceptual difficulty relative to T1 and T4. To explore error distributions in more detail, incorrect responses were modeled with a GLMM, and a tone confusion matrix was produced and visualized in the form of a heatmap (see Fig. 23 below).

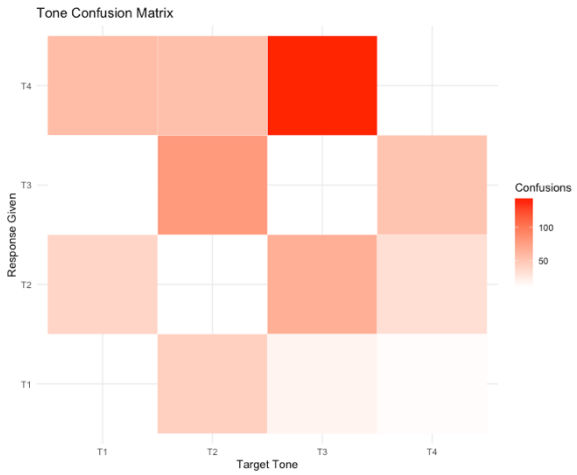


Figure 23 Disyllabic target identification test confusion matrix

As observable, the overall top confusions in disyllabic targets are the following:

T3 → T4 (n = 143)

T2 → T3 (n = 79)

T1 → T4 (n = 56)

To capture potentially complex interdependencies between the target tone, syllable position, and context tone, we first fit a GLMM with all main effects and their three-way interaction. The model included all main effects, two-way interactions, and the three-way *Condition.Sylpos.OtherTone*, with random intercepts for both Student and Item to account for repeated measures and item-specific variance.

Given the overparameterization and convergence failures of the three-way model, we estimated a more parsimonious focused interaction model, which tested the main effects of *Condition*, *Sylpos*, and *OtherTone*, the interaction between *Condition* and *Sylpos*, and included random intercepts for Speaker and Item.

Results of the model, with T1 as intercept, are summarized in Tab. 13 below:

Table 13 Disyllabic target identification test summary table

Term	Estimate	p-value	Interpretation
(Intercept)	+0.68	* p = .011	Baseline log-odds: T1 on first syllable, OtherTone = T1 (~66% prob).
ConditionT2	-1.00	*** p < .001	T2 significantly harder to identify than T1.
ConditionT3	-2.07	*** p < .001	T3 the hardest to identify.
ConditionT4	-0.71	** p = .007	T4 harder than T1 overall.
SylposSyl2	+0.49	. p = .074	Weak trend: second syllable slightly easier overall.
OtherToneT2/T3/T4	n.s.		No significant impact from the tone of the other syllable.
ConditionT4:SylposSyl2	+2.58	*** p < .001	Very strong facilitation: T4 becomes dramatically easier on second syllable.

The main findings of the tone identification task reveal several key patterns in learners' perceptual accuracy. First, robust tone category effects emerged, with T3 and, to a slightly lesser extent, T2 proving significantly more difficult to identify than T1 and T4. This is consistent with previous research identifying T3 as particularly challenging for L2 learners due to its complex, contour-based pitch shape and variable phonetic realization across contexts. T2

also presented difficulties, likely owing to its rising contour, which is prone to confusion with both T3 and T1 in non-native perception.

Interestingly, syllable position effects were particularly salient for T4. While T4 was quite difficult to identify when it occurred in the first syllable, its identification rate improved dramatically when it was positioned in the second syllable. This suggests that the prosodic context or boundary cues available at the end of the phrase may facilitate perception of T4's falling contour, which is otherwise susceptible to compression in non-final position (see § 2.2.1.4).

Finally, the variable OtherTone – which indexed the tone of the non-target syllable in disyllabic items – did not reach statistical significance. This indicates that participants were, on the whole, successful in isolating and identifying the tone of the target syllable, despite the presence of potential carryover effects (see § 2.3.3). In other words, learners were able to maintain attentional focus on the relevant syllabic unit, suggesting a level of tonal segmentation skill that supports more advanced perceptual strategies.

To visualize the interaction between identification condition and syllable position, we employed the *emmeans* package in R to extract and plot the estimated marginal means (EMMs) from the fitted model.

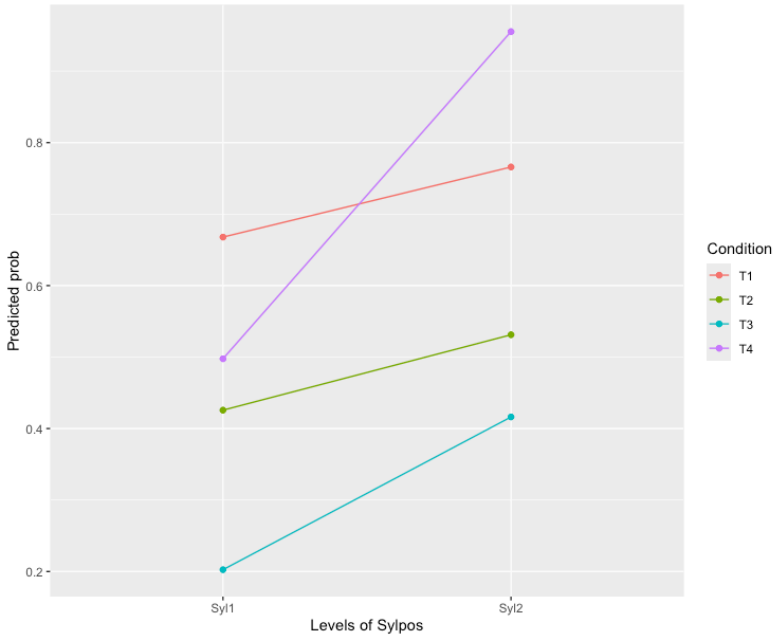


Figure 24 Tone identification accuracy by tone and syllable position

Taken together, these findings indicate that T1 is more readily identified than T4 when the target occurs on the first syllable, whereas T4 surpasses T1 in ease of identification when

situated on the second syllable, reflecting a robust positional facilitation effect specific to T4. Accordingly, the overall hierarchy of tonal identification difficulty in these disyllabic contexts can be summarized as follows:

T4≈T1>T2>T3

This pattern underscores the relative perceptual salience of T4 in phrase-final position and reinforces well-established challenges associated with the accurate perception of T2 and, most notably, T3 (Tu et al., 2016; Yang and Yang, 2017; Chen and Li, 2023; *inter alia*).

Overall identification score

An aggregate score for each student was derived by computing the arithmetic mean of their performance on the two identification tests (see Appendix E).

Oral competences assessment (HSKK test)

To assess the consistency between evaluators in rating the L2 Mandarin production data, we conducted inter-rater reliability analyses using the Intraclass Correlation Coefficient (ICC) in R. The ICC is particularly suited for evaluating agreement among raters on continuous scores and is preferred over simple correlation as it accounts for both systematic and random differences in scoring.

Two separate datasets were analyzed according to learners' affiliation. The first dataset (Affiliation=SI) consisted of 26 learners evaluated independently by two raters, hereby named Rater A and Rater B. The second dataset (Affiliation=RM) included 16 different learners, evaluated by the hereby named Rater B and Rater C evaluators, where Rater B served as the common rater across both datasets. Since the same rating criteria were followed, we adopted a two-way mixed-effects model assuming the raters are fixed and that we are interested in the consistency of their ratings. The analysis was conducted using the ICC() function from the psych package in R.

The results for the first dataset reported an ICC(3,1) = 0.488 (p = 0.0057), indicating moderate agreement between raters, and an ICC(3,k) = 0.656, suggesting a moderate reliability of the average score across both raters. Given this limited agreement, we identified five learners (S7, S15, S18, S22, S23) with particularly large discrepancies between raters (Rater A and Rater B). For these students, a new evaluator (Rater D) was involved, who provided overall scores averaged across the two subtasks (task1 and task2). The updated scores for these five

learners were recalculated by averaging the scores from all three evaluators, as reported in Table 14 below.

Table 14 HSKK test final agreement score for moderately reliable rates

Speaker	Rater A	Rater B	Rater D	Final Score
S7	20.5	9.0	16.5	15.33
S15	26.0	14.0	22.0	20.67
S18	21.0	9.0	18.0	16.00
S22	33.0	19.0	27.0	26.33
S23	27.5	10.0	20.5	19.33

In the second dataset, the ICC analysis revealed a substantially higher agreement: ICC(3,1) = 0.74 (p = 0.00031), with a corresponding ICC(3,k) = 0.85, indicating good to excellent inter-rater reliability. This result suggests that the scores from Rater B and Rater C can be reliably averaged to produce a single composite score per learner without requiring additional raters.

Appendix F provides the final scores from the entire dataset, including participants from both affiliation groups.

Overall proficiency score and clustering

In order to compute an overall proficiency score, data were first normalized by transforming both the identification and HSKK scores into z-scores. Subsequently, a Pearson correlation analysis was conducted to examine the relationship between the two measures.

The analysis yielded a correlation coefficient of r=0.35 with a p-value of 0.022 and a 95% confidence interval of [0.054, 0.593]. Given that the p-value is below the conventional threshold of 0.05, we conclude that the correlation is statistically significant, indicating that the two test scores are not independent. However, the magnitude of the correlation suggests only a moderate positive association. Thus, while there is sufficient alignment to support the combined interpretation of the two tests, the relationship is not so strong as to render them redundant.

We subsequently conducted a Principal Component Analysis (PCA) to assess whether a single underlying latent dimension could adequately account for the majority of the variance observed across the two test measures.

The PCA results indicate that the first principal component (PC1) accounts for 67.6% of the total variance. This suggests that a single latent "proficiency" axis – captured by PC1 – effectively summarizes most of the shared variance between the two tests. Nonetheless, approximately one-third of the variance remains unaccounted for by this dimension, indicating that while PC1 provides a meaningful composite index, it does not fully encapsulate all aspects of performance assessed by the two tests.

Table 15 PCA results for overall proficiency score

Component	Std. Deviation	Variance Explained	Cumulative Variance
PC1	1.163	67.6%	67.6%
PC2	0.805	32.4%	100%

Based on the outcomes of the PCA and the correlation analysis – both of which supported moderate shared variance yet retained sufficient distinctiveness – we proceeded to employ both z-scored standardized variables (*HSKK_z* and *Id_z*) as inputs for clustering participants into proficiency levels. Clustering was performed using the k-means algorithm, a partitional clustering method that partitions observations into *k* non-overlapping groups by minimizing the within-cluster sum of squares (WSS). This approach effectively groups individuals based on similarity in their standardized scores.

To determine an appropriate number of clusters, we applied the elbow method using within-cluster sum of squares (WSS). The elbow plot suggested that either two or three clusters could be justified. Consequently, we explored solutions for both k=2 and k=3.

Following clustering, we examined the centroids of each cluster to characterize proficiency levels across the standardized oral (HSKK) and identification (Id) measures. Specifically, we computed the mean scores within each cluster for both the two-cluster and three-cluster solutions. Clusters were then re-labeled post hoc to reflect an ascending proficiency continuum based on their centroid scores (see Appendix F).

To further assess whether a two- or three-cluster solution provided a more meaningful representation of underlying proficiency levels, we complemented visual inspection of cluster plots with a quantitative evaluation of clustering quality via silhouette analysis.

The results, summarized below, indicate that the two-cluster solution exhibited a higher average silhouette width, thus providing stronger support for adopting a two-level proficiency classification.

$k = 2$: Average silhouette width = 0.457

$k = 3$: Average silhouette width = 0.372

Given the the above-mentioned results and for parsimony reasons, the simpler k=2 model is marginally preferred and will be retained for subsequent analyses.

3.5.1.2 Musicality variable

Musical Sophistication Index

As outlined in the Methods section, the Goldsmiths Musical Sophistication Index (Gold-MSI) questionnaire was incorporated as a separate section within the background questionnaire to facilitate participants' self-assessment of musical sophistication (see § 3.4.1.2).

The Gold-MSI is a validated psychometric instrument designed to assess individual differences in musical abilities. It yields a General Musical Sophistication score, which represents a latent construct derived from a weighted combination of items drawn across five subscales: Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions. This score was computed using the official scoring template made available by the authors[13]. Higher scores indicate greater musical sophistication relative to the population average. In our cohort, descriptive statistics for the general factor and subscale scores are reported in Appendix G.

Tone-deafness test

A hyperlink to the tone-deafness test was incorporated into a dedicated section of the background questionnaire. Participants were instructed to complete the test in a quiet environment, adhering closely to the guidelines provided on the official test website[14]. Upon completing the assessment, participants were instructed to submit their individual score plots to the researcher. The tone-deafness test yields scores ranging from fewer than 18 to a maximum of 33. The raw scores are provided in Appendix G.

Overall musicality score

To derive an overall musicality score, we incorporated results from the two above-mentioned established measures of musical aptitude: the Gold-MSI and the tone-deafness test.

[13] Available at https://osf.io/9ytz2/ (Accessed Sept 22, 2025)
[14] Available at https://www.themusiclab.org/quizzes/td (Accessed Sept 22, 2025)

While both instruments are theorized to assess facets of musical ability, they capture distinct domains: the Gold-MSI primarily reflects broader experiential and behavioral engagement with music, whereas the tone-deafness test specifically indexes perceptual acuity in pitch discrimination. Since both perceptual and experiential components of musicality have been hypothesized to facilitate the learning and production of lexical tones, we analyzed these measures jointly to obtain a more integrated characterization of participants' musical aptitude.

Across the sample of 42 participants, raw MSI scores ranged from 38 to 109, whereas tone-deafness scores spanned from 17 to 33. Preliminary correlation analyses indicated a modest, non-significant positive relationship between MSI and tone-deafness scores (Pearson's $r=0.16$, $p=0.31$; Spearman's $\rho=0.20$, $p=0.21$). This weak correlation suggests that, although conceptually related, these measures likely capture complementary and partially independent dimensions of musicality. To facilitate their direct combination despite differing scales, both scores were subsequently standardized (z-transformed).

We constructed a composite musicality score employing two complementary approaches:

1. Z-Mean Method

For each participant, we computed the unweighted average of their standardized scores on the MSI and tone-deafness test measures. This approach assumes that experiential (MSI) and perceptual (tone-deafness) components contribute equally to overall musicality.

2. Principal Component Analysis (PCA)

We conducted an exploratory PCA on the two z-scored variables to derive a data-driven latent musicality factor. The first principal component (PC1) displayed equal positive loadings from both the MSI and the tone-deafness test scores (loadings = 0.762), accounting for 58.1% of the total shared variance between the two measures.

A summary table containing the relevant loadings and variance explained is provided in Appendix G.

Because both variables contributed almost equally to the first principal component, and to preserve the conceptual clarity afforded by equal-weighted z-averaging, we adopted the z-mean composite score for use in the subsequent models. This decision was further motivated by considerations of interpretability and theoretical transparency, particularly given the limited number of input variables.

To categorize participants into meaningful groups based on their overall musical ability, we applied an unsupervised clustering procedure to their composite musicality scores. This process aimed to reduce dimensionality, simplify interpretation, and enable comparisons across levels of musicality. To classify participants into discrete musicality groups, we applied k-means

clustering to their Musicality_zmean scores. This analysis was facilitated by *stats::kmeans()* for clustering, *cluster::silhouette()* for internal validation, and supplemented by *dplyr* and *ggplot2* for data processing and visualization.

To determine the optimal number of clusters, we computed average silhouette widths for solutions with two and three clusters. Results indicated:

$k = 2$: Average silhouette width = 0.590

$k = 3$: Average silhouette width = 0.565

Although both solutions suggested meaningful structure, the two-cluster model exhibited marginally better cohesion and separation. For conceptual clarity and interpretability, we therefore proceeded with the two-cluster solution, distinguishing participants as either *low* or *high* in musicality (see Appendix G).

3.5.2 Pre-modelling variable screening

A correlation analysis was conducted to quantify the degree of statistical association among key variables, serving as an initial screening step to inform subsequent model specification. This procedure facilitated the identification of predictors that were sufficiently independent to justify simultaneous inclusion in downstream analyses. Specifically, the analysis examined associations between:

- University year (hereinafter Grade);
- Proficiency level, based on the two-level k-means solution;
- Musicality scores, based on the two-level k-means solution.

For the computation of the correlation matrix, all variables were consolidated into a single dataset to enable pairwise correlation estimates. This approach ensured that potential multicollinearity could be systematically evaluated prior to formal modeling. The results of the analysis are summarized in the following table:

Table 16 Pairwise correlation estimates (variable screening)

Variable Pair	**Pearson's *r***	**Interpretation**
Proficiency ~ Grade	+0.07	No linear relationship between proficiency cluster and university year

Variable Pair	Pearson's *r*	Interpretation
Proficiency ~ Musicality	+0.28	Modest positive relationship
Grade ~ Musicality	-0.12	No correlation

The near-zero correlation between cluster-based Proficiency and Grade underscores that curricular progression does not necessarily align with phonetic proficiency. This result emphasizes the importance of incorporating direct, performance-based assessments of proficiency, thereby justifying their role as primary predictors in subsequent analyses.

Finally, the modest positive correlation between Musicality and Proficiency suggests that individual differences in auditory skills may relate with phonetic acquisition; however, the relatively weak strength of this association supports treating Musicality and Proficiency as complementary yet non-redundant predictors in the planned statistical models.

4. L2 Mandarin Tone production in isolated target words

The following control study investigates the tone production of monosyllabic and disyllabic target phrases uttered in isolation by intermediate Italian university learners of Mandarin (n = 42). The primary objective is to model the fundamental frequency (F0) trajectories characterizing each tone, while rigorously accounting for inter-speaker variability.

4.1 Research questions and hypotheses

This component of the project pursues two methodological goals: (i) to deliver a baseline of tone-specific F0 trajectories (T1-T4) for Italian intermediate learners in controlled monosyllabic and disyllabic productions, and (ii) to establish reference contours against which tone production in disyllabic targets embedded in intonational phrases can be systematically compared in subsequent analyses. Hence, the research questions (RQs) are as follows:

RQ1. What are the F0 contours of each lexical tone (T1-T4) when learners produce monosyllables and disyllables in isolation?

RQ2. To what extent do learners' tone realizations in monosyllabic contexts approximate canonical citation forms in both pitch register (F0 height) and contour shape (time-varying trajectory)?

Given the cohort's intermediate proficiency, we advance the following hyphotesis (H):

H1. In monosyllabic targets, learners' tone realizations will broadly approximate canonical citation forms, both in pitch height and contour shape, with no large or systematic deviations from normative targets.

H2. Accuracy will be slightly reduced in disyllabic targets due to inter-syllabic tonal coarticulation and positional asymmetries, yielding greater variability in F0 scaling and contouring than in monosyllables, but still not affecting general tone phonological accuracy.

H3. Because the phrase-final syllable (Syl2) may benefit from boundary-related stabilization, we tentatively predict higher accuracy on Syl2 than on Syl1 in producing tones in their citation forms. This remains an empirical question, as phrase-final position can also introduce compression or boundary effects that could counteract such an advantage.

4.2 Dataset Overview

Two complementary datasets were compiled to model tone production in monosyllabic and disyllabic targets. Both corpora were annotated at ten time-normalized points per syllable and include speaker-normalized F0 (z-scores) to enable cross-speaker comparisons.

The monosyllabic dataset includes 6,720 observations. For each token, the following variables were annotated:

- Speaker: Individual speaker identifier;
- Grade, Proficiency, Musicality: speaker-level categorical predictors indexing academic level, L2 proficiency, and musical aptitude;
- Tone: lexical tone category (T1-T4);
- Item: four different segmental items per tone;
- Point: time-normalized sampling point (1-10) across the syllable;
- F0: raw fundamental frequency (Hz);
- F0_z: speaker-normalized F0 (z-score).

The disyllabic dataset includes 13,420 observations with parallel annotation, enriched for positional and coarticulatory context:

- Speaker: unique speaker identifier;
- Grade, Proficiency, Musicality: speaker-level categorical predictors (as above);
- SylPos: syllable position within the disyllable (Syl1, Syl2);
- Tone: lexical tone of the target syllable (T1-T4);
- OtherTone: lexical tone of the adjacent syllable (preceding or following, depending on SylPos);
- Point: time-normalized sampling point (1-10) across the syllable;
- F0: raw fundamental frequency (Hz);
- F0_z: speaker-normalized F0 (z-score).

4.3 Monosyllabic target

This section reports on the analysis of learners' production of Mandarin lexical tones in monosyllabic contexts, aimed at establishing a baseline for tonal accuracy prior to examining prosodic integration in disyllabic phrases.

4.3.1 Establishing the Baseline Model for monosyllabic Tone production Contours

A Generalized Additive Mixed Model (GAMM) was fit using R package *mgcv* to estimate smooth F0 trajectories over time for each tone category. Random smooth terms of *Point* were included by *Speaker* and *Item* to account for speaker- and item-level variability. The function *bam()* was chosen instead of *gam()* because it is optimized for large datasets and allows efficient estimation of complex models with random effects. The model converged successfully (fREML = 7118.4), with an estimated scale parameter of 0.465 and adjusted R^2 = 0.532, explaining approximately 54.3% of the deviance. Model diagnostics (*gam.check*) confirmed that the basis dimension (k = 10) was appropriate for all smooth terms (k-index ≈ 0.98, all p > .10), indicating no signs of undersmoothing.

The population-level smooths are visualized in the resulting contour plots, reflecting fixed effects of Tone on F0 trajectories without incorporating individual deviations. The intercept corresponds to T1, serving as the reference category. All other tones revealed significantly lower mean pitch levels:

Table 17 Tone mean pitch levels compared to T1 (intercept)

Tone	Estimate	t-value	p-value
T1	0.85	NA	NA
T2	-1.24	-13.16	< .001 ***
T3	-1.36	-42.36	< .001 ***
T4	-0.69	-13.30	< .001 ***

Hence, the rank order of pitch height in monosyllabic target production aligns with phonological expectations:

T1 > T4 > T2 > T3

Analysis of smooth terms across time revealed that T2, T3, and T4 displayed significant non-linear pitch movements, consistent with their canonical shapes – rising (T2), dipping (T3), and falling (T4). In contrast, T1 maintained a flat, high-level trajectory. The smooth terms are listed in Tab. 18 and visualized in Fig. 25.

Table 18 Tone smooth terms values

Smooth Term	edf	F	p-value
s(Point):ToneT1	2.32	2.18	0.114
s(Point):ToneT2	5.01	41.61	< .001 ***
s(Point):ToneT3	5.47	62.28	< .001 ***
s(Point):ToneT4	3.96	33.25	< .001 ***

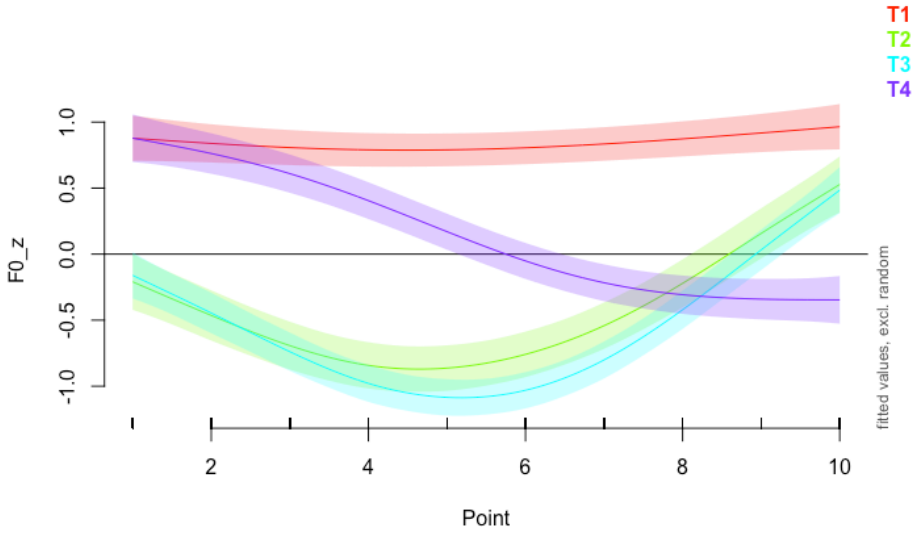


Figure 25 GAMM curves by Tone in monosyllabic productions

Given their similarity in the GAMM plot, a pairwise difference smooth analysis was conducted to further examine distinctions between T2 and T3. Both tones revealed internal variability ($p < .001$), but significant differences between their trajectories emerged only in a localized region – specifically between time points 5.5-7.0. These findings suggest broad contour similarity, punctuated by subtle pitch distinctions in select temporal segments, pointing to partial overlap in learner production of rising (T2) and dipping (T3) tones.

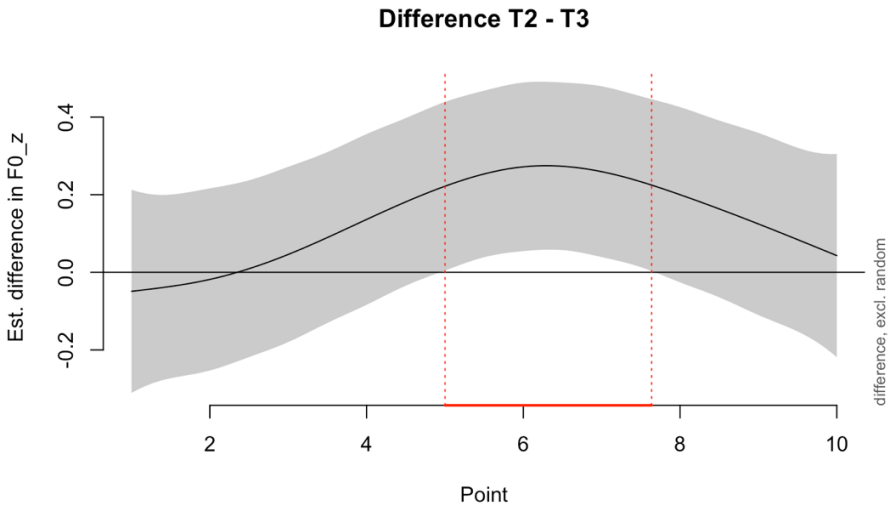


Figure 26 T2 vs T3 pairwise difference smooth in monosyllabic productions

4.3.2 Evaluating the Influence of Learner Factors Against the Baseline Model

To investigate the extent to which individual-level factors influenced tonal realization, the baseline model was extended to include interaction terms between Tone and three speaker-level predictors: Proficiency, Musicality, and Grade. Each extended model was evaluated against the baseline using maximum likelihood (ML) estimation and likelihood ratio tests. All three models provided statistically significant improvements over the baseline:

Table 19 Proficiency, Musicality and Grade models comparison with the baseline

Model	ΔAIC	χ^2(df)	p-value
Tone.Proficiency	1867.3	$\chi^2(15) = 939.15$	< .001 ***
Tone.Musicality	1708.9	$\chi^2(15) = 858.79$	< .001 ***
Tone.Grade	1833.5	$\chi^2(27) = 922.98$	< .001 ***

The Tone.Proficiency model yielded a substantial improvement in fit (ΔAIC = 1867.3; $\chi^2(15) = 939.15$), indicating that the effect of Tone on F0 trajectories varies meaningfully with L2 proficiency level. Similarly, the Tone.Musicality model also demonstrated a significant improvement (ΔAIC = 1708.9; $\chi^2(15) = 858.79$), suggesting that musical aptitude modulates tonal production patterns. Finally, the Tone.Grade model provided the largest improvement (ΔAIC = 1833.5; $\chi^2(27) = 922.98$), implying that academic level is a relevant factor in monosyllabic tonal differentiation in isolation.

For interpretability and visualization, all models were refitted using REML method.

4.3.3 Interaction Between Proficiency and Tone production

The model Proficiency.Tone included interaction-specific smooth terms over time (*Point*) for each Proficiency.Tone combination, as well as random smooths for Speaker and Item[15].

To examine whether and how Proficiency modulates the realization of Mandarin lexical tones, we conducted pairwise comparisons of estimated marginal means (EMMs) for normalized F0 (i.e., F0_z) between high- and low-proficiency groups within each tone, based on a GAMM.

Tab. 20 summarizes the estimated differences in F0_z between high- and low-proficiency speakers for each tone:

Table 20 Estimated differences in F0_z between high- and low-proficiency speakers

Tone	Estimate (High-Low)	SE	t-ratio	p-value	Significance
T1	-0.0520	0.048	-1.07	0.28	n.s.
T2	-0.0564	0.082	-0.69	0.49	n.s.
T3	-0.1452	0.086	-1.69	0.09	marginal (n.s.)
T4	+0.5469	0.075	7.27	<.0001	significant

For T1,T2, and T3, no statistically significant differences were observed in F0_z between high- and low-proficiency learners ($p > .05$). This suggests that L2 proficiency has a limited impact on the pitch realization of these tones in monosyllabic productions. In contrast, a highly significant difference was found for T4 ($p < .0001$), with high-proficiency learners producing substantially higher F0_z values than their low-proficiency counterparts. As illustrated in Fig. 27, this difference is concentrated primarily at tone onset, where high-proficiency learners (red line) exhibit a steeper and more target-like pitch trajectory resulting from a higher initial F0_z.

[15] The model converged successfully with no diagnostic issues, explaining 55.9% of the deviance (adjusted R^2 = 0.547). Basis dimension checks confirmed that the chosen k = 10 was adequate for all smooth terms (k-index = 0.99, all p > .20).

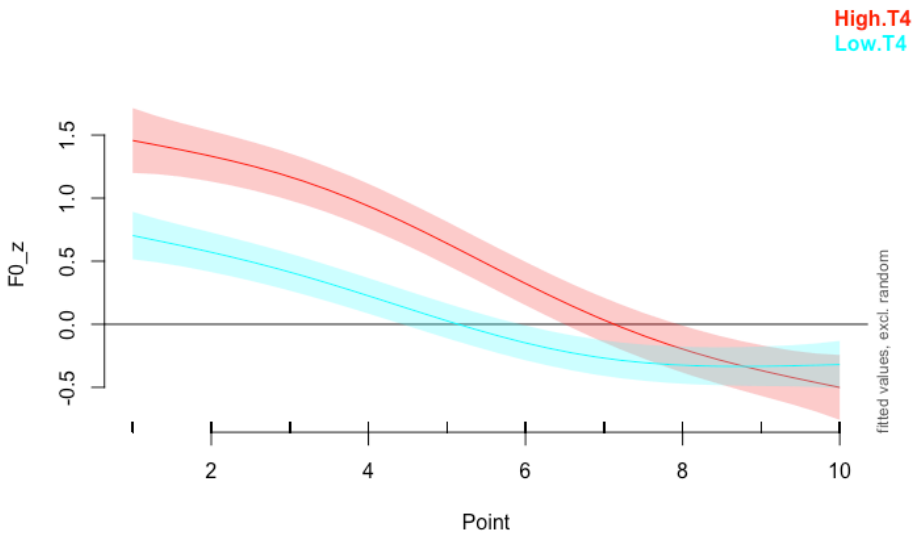


Figure 27 T4 by Proficiency in monosyllabic productions

4.3.4 Interaction Between Musicality and Tone production

To assess whether musical aptitude influences the production of Mandarin lexical tones, we conducted pairwise comparisons of EMMs for F0_z between high- and low-musicality learners within each tone, based on a GAMM with smooths over time for each Musicality.Tone combination.

Tab. 21 summarizes the estimated F0_z differences (High-Low musicality) for each tone:

Table 21 Estimated F0_z differences between high- and low-musicality speakers

Tone	Estimate (High-Low)	SE	t-ratio	p-value	Significance
T1	-0.0864	0.054	-1.58	0.11	n.s.
T2	-0.0780	0.077	-1.01	0.31	n.s.
T3	+0.2694	0.081	3.31	0.0009	significant
T4	-0.1782	0.064	-2.79	0.0052	significant

For T1 and T2, there were no statistically significant differences in F0_z between high- and low-musicality learners ($p > .05$), suggesting that musical aptitude does not substantially impact the realization of these tones in monosyllabic contexts. For T3, high-musicality participants produced higher F0_z values than their low-musicality peers. However, as we can observe from the GAMM plot (Fig. 28) such significance lays mainly in the central dipping portion of the contour.

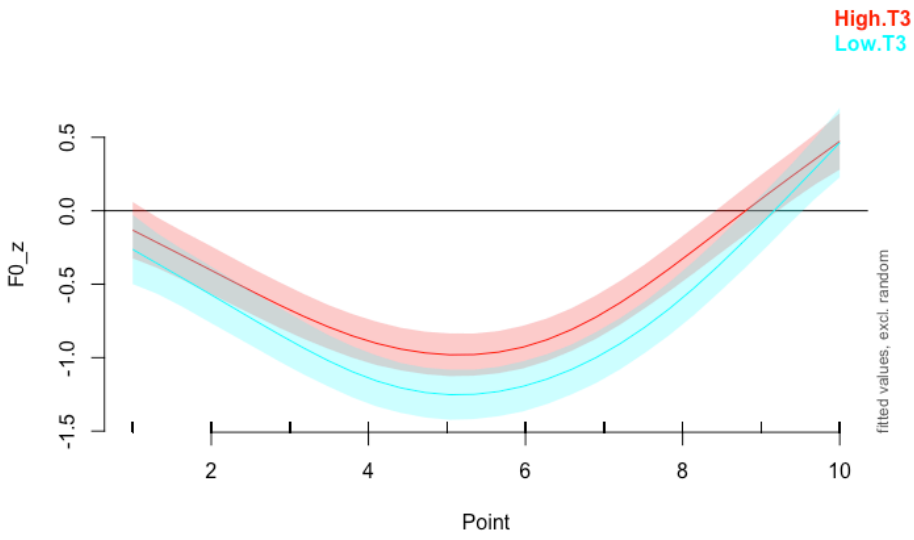


Figure 28 T3 by Musicality in monosyllabic productions

For T4, a significant effect was also found, but in the opposite direction: high-musicality participants produced lower F0_z values compared to the low-musicality group. As we can observe from the GAMM plot (Fig.29), the significance lays mainly on the second half of the portion, where the contour flatters. This flattening may in fact be the result of a creakiness, which is common in the phonetic realization of T4 final portion, also in native speakers.

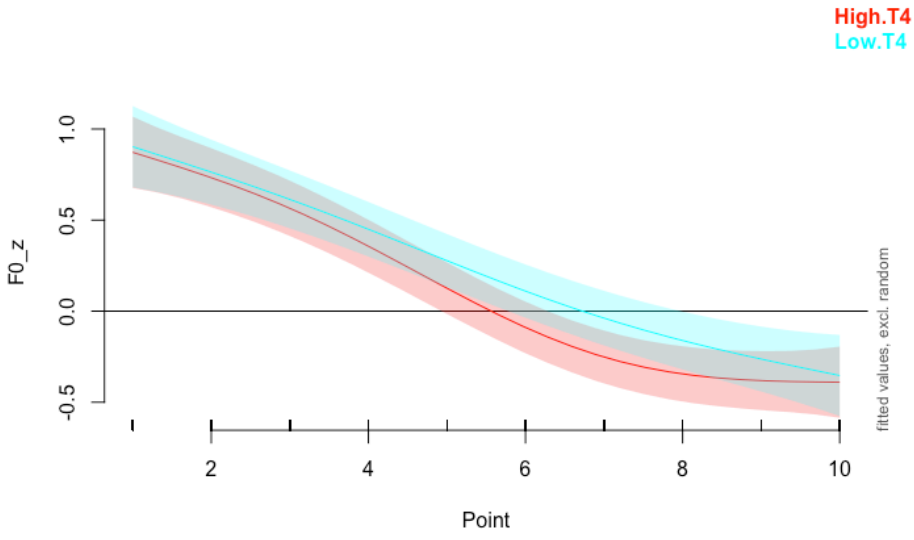


Figure 29 T4 by Musicality in monosyllabic productions

4.3.5 Interaction Between Academic Level and Tone production

To investigate the influence of academic progression on Mandarin tone production, we compared F0_z across three grade levels – BA2 (bachelor second year), BA3 (bachelor third year), and MA1 (master first year) – within each tone category. EMMs and pairwise contrasts were derived from a GAMM with smooths over time for each Grade.Tone combination.

Tab. 22 summarizes the estimated differences in F0_z between grade levels for each tone, with p-values adjusted by the Tukey method.

Table 22 Estimated differences in F0_z between grade levels

Tone	Contrast	Estimate	SE	t-ratio	p-value	Significance
T1	BA2 - BA3	0.1614	0.0532	3.04	0.0068	significant
T1	BA2 - MA1	-0.0633	0.0432	-1.47	0.3073	n.s.
T1	BA3 - MA1	-0.2246	0.0511	-4.39	<.0001	significant
T2	BA2 - BA3	-0.1041	0.0875	-1.19	0.4588	n.s.
T2	BA2 - MA1	0.0366	0.0863	0.42	0.9055	n.s.
T2	BA3 - MA1	0.1408	0.0862	1.63	0.2315	n.s.
T3	BA2 - BA3	-0.1455	0.0904	-1.61	0.2418	n.s.
T3	BA2 - MA1	0.0840	0.0889	0.95	0.6116	n.s.
T3	BA3 - MA1	0.2295	0.0896	2.56	0.0282	*significant*
T4	BA2 - BA3	0.2257	0.0615	3.67	0.0007	significant
T4	BA2 - MA1	-0.1466	0.0630	-2.33	0.0525	. (marginal)
T4	BA3 - MA1	-0.3723	0.0749	-4.97	<.0001	significant

For T1, significant differences were observed between BA2 and BA3 (p = .0068) and between BA3 and MA1 (p < .0001), with BA3 speakers producing lower F0_z than both BA2 and MA1. No significant difference was found between BA2 and MA1.

For T2, no statistically significant differences emerged between any pairwise comparisons across academic levels (Grade), suggesting a consistent realization of this tone among learners regardless of university year.

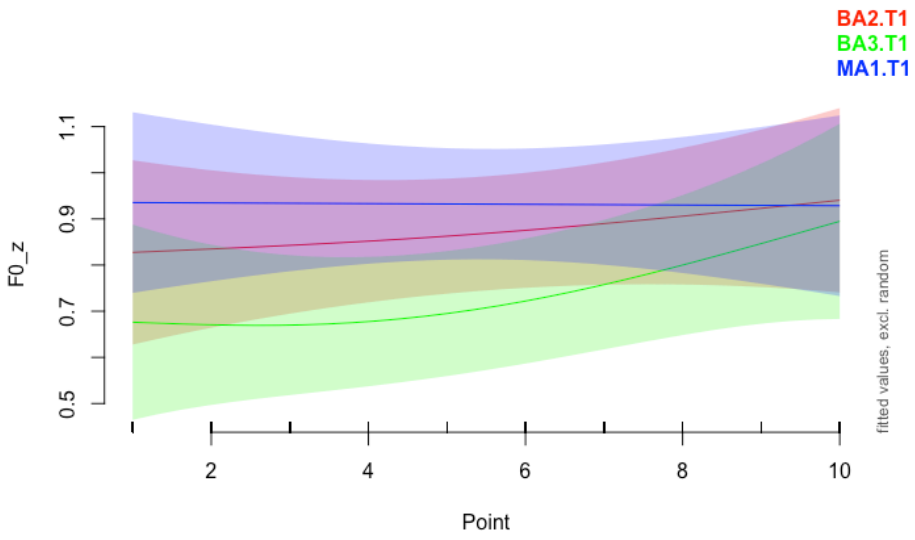


Figure 30 T1 by Grade in monosyllabic productions

In contrast, T3 exhibited a significant difference between BA3 and MA1 students ($p = .0282$), with BA3 learners producing systematically higher F0_z values. All other pairwise comparisons for T3 were non-significant.

Inspection of the predicted F0 contours reveals that this difference is localized primarily in the second half of the syllable, where BA3 productions consistently display a higher pitch trajectory relative to those of MA1 learners.

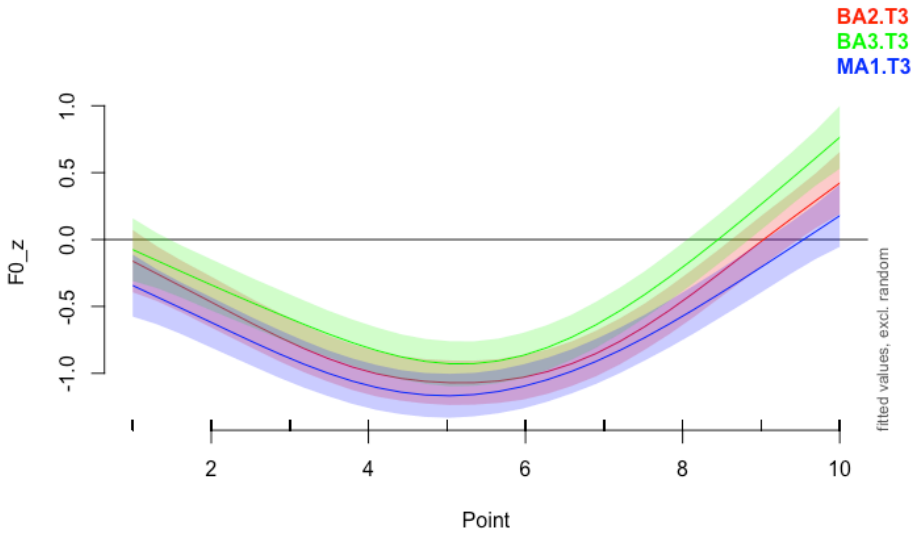


Figure 31 T3 by Grade in monosyllabic productions

For T4, significant differences in F0_z values were observed across academic levels. In particular, BA3 learners produced significantly lower F0_z than both BA2 and MA1 groups. The contrast between BA2 and MA1 approached significance ($p = .0525$), indicating a potential but still marginal trend.

Visual inspection of the GAMM-derived F0 contours (see Fig. 32) reveals that the primary divergence occurs in the initial portion of the syllable, where MA1 learners exhibit a steeper and higher onset compared to both BA groups. This pattern suggests that MA1 students are more successful in initiating the canonical falling contour of T4, resulting in a more native-like convex trajectory. In contrast, the flatter initial slope observed in BA learners may reflect a reduced ability to execute the rapid pitch descent characteristic of this tone.

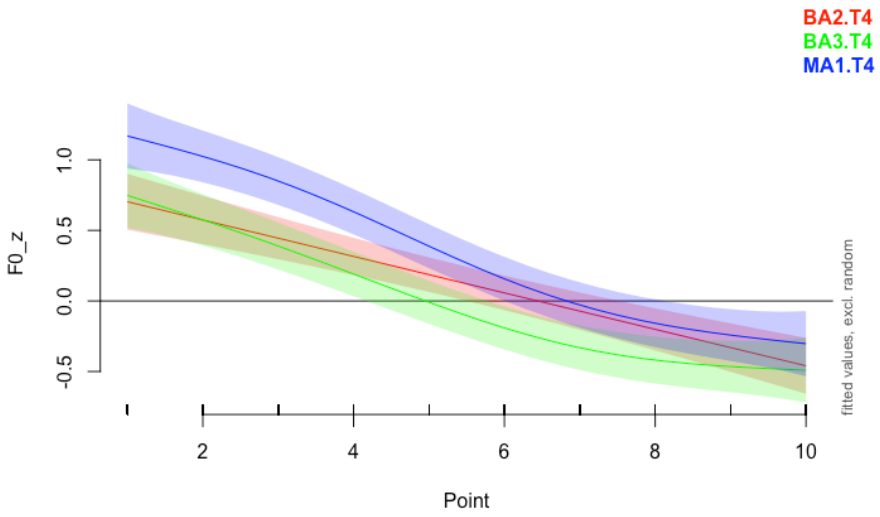


Figure 32 T4 by Grade in monosyllabic productions

4.3.6 Interim summary on monosyllabic targets

This section examined the realization of Mandarin lexical tones in monosyllabic contexts among Italian learners of Mandarin, with a focus on the influence of tone category, L2 proficiency, musical aptitude, and academic level on pitch contour realization.

Results from the parametric terms pointed to a pitch height hierarchy of T1 > T4 > T2 > T3, consistent with phonological descriptions of Mandarin tones.

Visual inspection of the GAMM-predicted trajectories indicated notable similarities between T2 and T3 productions. A pairwise smooth comparison revealed only minimal differences across the temporal domain, with statistically significant divergence restricted to a brief, localized interval in the central portion of the contour. These findings suggest that

learners' productions of T2 and T3 largely converge in overall shape, which may reflect an incomplete phonological contrast in L2 tonal categories. In particular, the reduced distinction at the syllable onset – where T2 is expected to be realized with a higher pitch than T3 – points to a potential area of category merger or insufficient target realization.

To assess how individual differences modulate tone production, three extended models were tested, incorporating interaction terms between Tone and: (i) Proficiency, (ii) Musicality, and (iii) Grade (academic level).

i. The Proficiency.Tone model revealed a robust effect for T4: high-proficiency learners produced significantly higher normalized F0 values compared to low-proficiency learners. This difference – primarily located in the initial portion of the contour – suggests improved control over the falling pitch movement associated with this tone. T1 and T2 revealed no significant group differences, while T3 showed only a marginal trend.

ii. The Musicality.Tone model revealed that musical aptitude influenced the realization of T3 and T4. High-musicality participants produced higher F0_z values for T3, particularly in the central portion of the contour. For T4, however, the effect was reversed: high-musicality learners produced lower F0_z in the final portion of the syllable, likely consistent with more accurate realization of final creaky phonation.

iii. The Grade.Tone model often highlighted deviations by BA3 learners, compared to BA2 and MA1 learners who revealed more similar trends. For T1, BA3 students produced lower F0_z values than both BA2 and MA1 students. T3 highlighted a significant difference between BA3 compared to MA1 and BA2 learners, with the latter producing lower F0_z values in the central dipping portion. The clearest developmental trajectory, where MA1 outperformed BA learners emerged again for T4: MA1 students produced higher F0_z values on the onset, resulting in a more convex shape and aligning more closely with target-like falling tone realizations.

While tone category remains the dominant predictor of F0 variation, significant modulations emerge based on L2 proficiency, musical aptitude, and academic level. T4 consistently emerged as the most sensitive to individual differences. Musical aptitude facilitated enhanced dynamic control, particularly for T3 and T4. Academic level tracked broader developmental shifts, with MA1 (and to some extent BA2) learners demonstrating more refined and target-like tone realizations.

4.4 Disyllabic target

This section investigates how Mandarin lexical tone and syllable position interact to shape pitch trajectories in disyllabic phrase production among Italian learners of Mandarin.

4.4.1 Establishing the Baseline Model for disyllabic Tone production Contours

To assess how tone category and syllable position jointly influence F0 contours, a GAMM was fitted with by-smooths over time (Point) for each Tone.Syllable Position combination, and random smooths for Speaker and OtherTone were included to account for inter-speaker and contextual tonal variability[16].

All parametric contrasts relative to the reference level (T1.Syl1) were highly significant ($p < .001$), reflecting robust differences in mean F0_z across both tone and syllable position. Notably, T2, T3, and T4 in Syl1 all showed substantially lower F0_z compared to T1.Syl1.

For the Syl2, all tones had further lowered mean F0_z relative to T1.Syl1, with T2.Syl2 and T3.Syl2 producing the lowest values overall.

For Syllable 1, T1 did not reveal a significant time-varying trajectory, aligning with its high-level pitch target; while T2, T3, and T4 exhibited significant non-linear F0 trajectories across time ($p < .001$). T2 and T3 contours mainly differed on the first portion of the predicted curve, with T2 starting higher; however on the final portion the predicted curves of T2 and T3 aligned in a slight rise. For Syllable 2, all four tones exhibited significant dynamic pitch trajectories, however only moderately significant for T1 ($p < .05$). Notably, on Syl2 all tones, T1 included, demonstrated more complex, time-varying contours, indicative of the impact of phrase-final position on tone production.

[16] The model converged successfully, with no indication of undersmoothing (all k-index ≈ 1.02, all $p > .89$). It explained 28.5% of the deviance (adjusted $R^2 = 0.275$), with a scale estimate of 0.75.

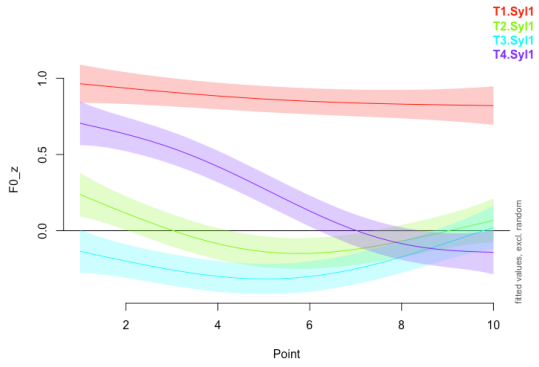


Figure 33 Tone production on syllable 1 in disyllabic targets

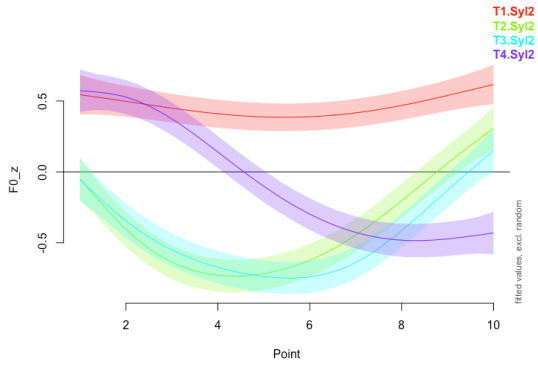


Figure 34 Tone production on syllable 2 in disyllabic targets

4.4.2 Evaluating the Influence of Learner Factors Against the Baseline Model

To determine whether individual-level factors such as proficiency, musicality, and academic grade improve the modeling of tonal F0 trajectories beyond tone and syllable position alone, we compared four GAMMs fit with maximum likelihood (ML) estimation and Akaike Information Criterion (AIC). Results are summarized in Tab. 23 below:

Table 23 Proficiency, Musicality and Grade models comparison with the baseline

Comparison	AIC Difference	Lower AIC Model	Interpretation
Base (TS) vs. Proficiency (PTS)	+112.24	Proficiency (PTS)	Adding proficiency substantially improves model fit.
Base (TS) vs. Musicality (MTS)	+82.90	Musicality (MTS)	Adding musicality also improves fit, but less than proficiency.
Base (TS) vs. Grade (GTS)	-11.19	Base (TS)	Including grade does not improve fit; in fact, the base model fits slightly better.

The addition of Proficiency yielded the greatest improvement in model fit, highlighting it as the most informative learner-specific predictor of tonal variation in disyllabic contexts. Musicality also accounted for a significant portion of the variance, suggesting that pitch processing and musical training may support more accurate tone implementation in dysillabic targets.

In contrast, academic grade level did not improve model fit and was even detrimental to parsimony. This finding suggests that formal academic progression is a less sensitive indicator of tonal performance in disyllabic phrase production in isolation.

4.4.3 Interaction between Proficiency, Tone production, and Syllable Position

To investigate how learner proficiency modulates the realization of individual Mandarin tones across syllable positions, we refitted the Proficiency.Tone.Syllable Position model using restricted maximum likelihood estimation (method = "fREML"). We then conducted pairwise comparisons of EMMs between high and low proficiency groups for each Tone.Syllable Position combination. The contrasts are summarized in the table below:

Table 24 Comparisons between high and low proficiency groups for each Tone.Syllable Position combination

Tone	Syllable	High-Low Estimate	SE	t	p-value	Significance
T1	Syl1	-0.115	0.084	-1.37	0.170	n.s.
T2	Syl1	-0.167	0.099	-1.69	0.092	n.s. (trend)
T3	Syl1	-0.307	0.103	-2.97	0.0029	significant
T4	Syl1	0.042	0.089	0.47	0.636	n.s.
T1	Syl2	0.027	0.090	0.30	0.763	n.s.
T2	Syl2	-0.124	0.117	-1.06	0.287	n.s.
T3	Syl2	0.361	0.112	3.21	0.0013	significant
T4	Syl2	0.552	0.116	4.77	<.0001	significant

For Syllable 1, only T3 showed a significant effect of Proficiency: High proficiency speakers produced substantially lower F0_z for T3 in the first syllable compared to low proficiency speakers (*estimate = -0.307, p = 0.0029*). For T1, T2, and T4, Proficiency did not yield significant differences, though T2 displayed a marginal trend (*p = 0.092*).

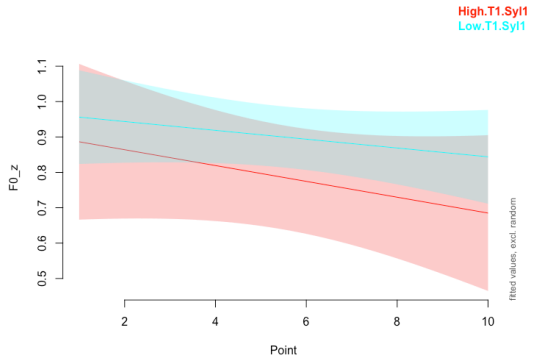


Figure 35 Tone 1 production on syllable 1 by Proficiency

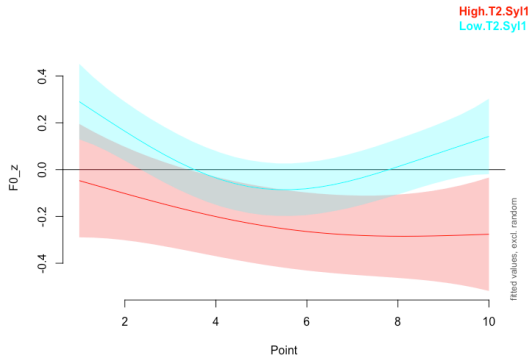


Figure 36 Tone 2 production on syllable 1 by Proficiency

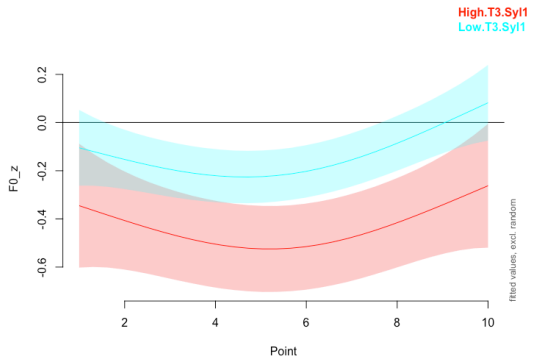


Figure 37 Tone 3 production on syllable 1 by Proficiency

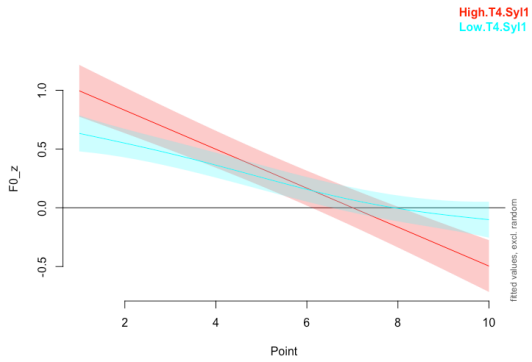


Figure 38 Tone 4 production on syllable 1 by Proficiency

For Syllable 2, T3 and T4 both exhibited significant proficiency effects, but with opposite directions: for T3, high proficiency speakers produced higher F0_z than low proficiency speakers. For T4, the difference was even more pronounced: high-proficiency participants produced substantially higher F0_z than low-proficiency speakers (estimate = +0.552, $p < .0001$), particularly in the initial segment, yielding a more convex and phonetically accurate falling contour. No significant Proficiency effects were found for T1 or T2 in the second syllable.

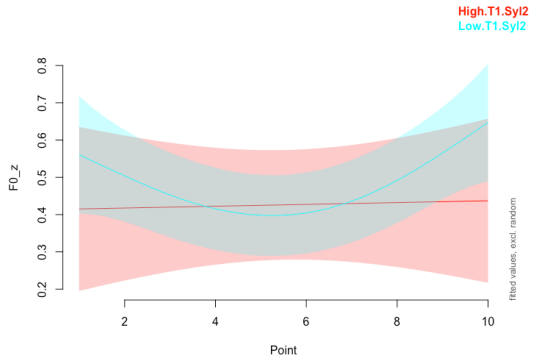


Figure 39 Tone 1 production on syllable 2 by Proficiency

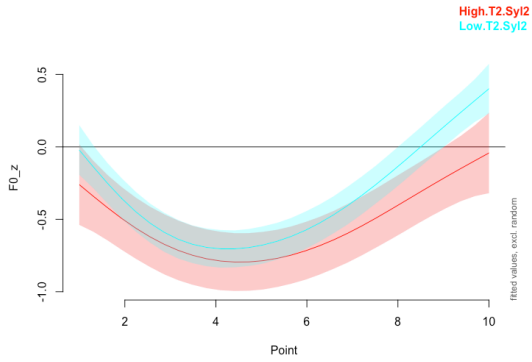


Figure 40 Tone 2 production on syllable 2 by Proficiency

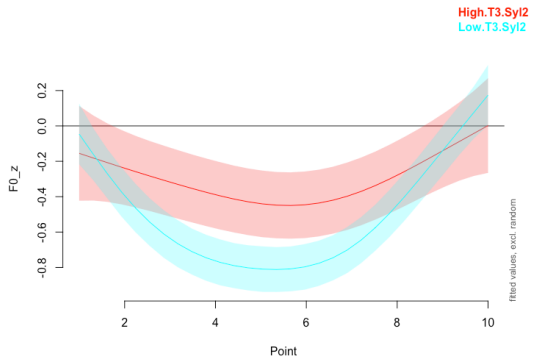


Figure 41 Tone 3 production on syllable 2 by Proficiency

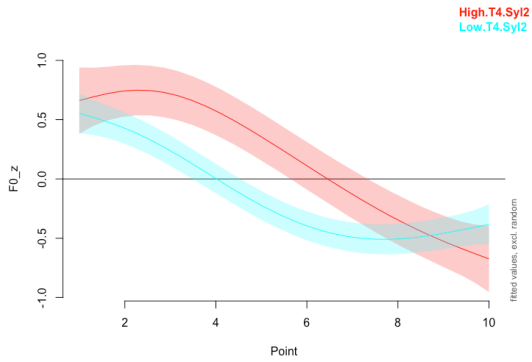


Figure 42 Tone 4 production on syllable 2 by Proficiency

These findings indicate that proficiency-based differences are tone-specific and syllable-position dependent. In particular, T3 and T4, again, indicated the strongest Proficiency effects across both syllables. The directionality of F0_z differences suggests that advanced learners are not only more consistent in differentiating tonal categories, but also in adjusting pitch targets in response to syllable position. Conversely, T1 and T2 exhibit greater stability across proficiency levels.

4.4.4 Interaction between Musicality, Tone production, and Syllable Position

To evaluate the impact of musical aptitude on the production of Mandarin lexical tones, we conducted pairwise comparisons of EMMs between High and Low musicality groups for each Tone.Syllable Position combination. The results are summarized in the table below:

Table 25 Comparisons between High and Low musicality groups for each Tone.Syllable Position combination

Tone	Syllable	High-Low Estimate	SE	t	p-value	Significance
T1	Syl1	-0.171	0.083	-2.06	0.040	*significant*
T2	Syl1	-0.092	0.098	-0.94	0.346	n.s.
T3	Syl1	-0.131	0.094	-1.40	0.162	n.s.
T4	Syl1	-0.169	0.084	-2.02	0.043	*significant*
T1	Syl2	0.152	0.094	1.61	0.107	n.s.
T2	Syl2	-0.173	0.109	-1.58	0.115	n.s.
T3	Syl2	0.382	0.110	3.49	0.0005	significant
T4	Syl2	0.316	0.106	2.99	0.0028	significant

For the first syllable (Syl1), significant effects of musicality were found for T1 and T4, with High musicality participants producing lower F0_z than their Low musicality peers. For T4, this lowering effect among high-musicality learners occurs primarily in the second portion of the predicted contour (see Fig. 46). No significant differences were found for T2 or T3.

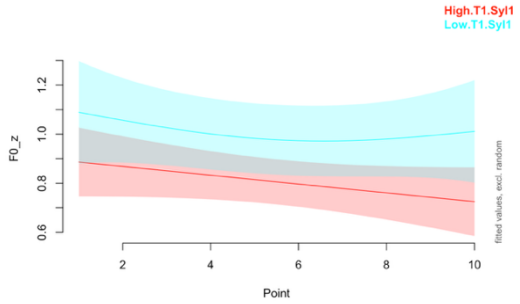


Figure 43 Tone 1 production on syllable 1 by Musicality

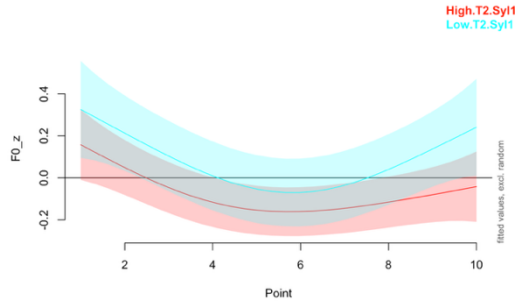


Figure 44 Tone 2 production on syllable 1 by Musicality

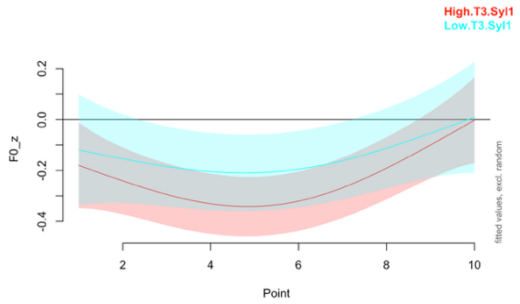


Figure 45 Tone 3 production on syllable 1 by Musicality

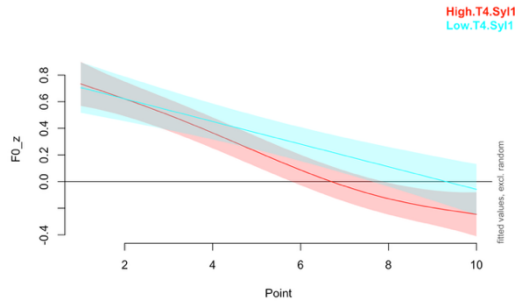


Figure 46 Tone 4 production on syllable 1 by Musicality

In Syl2 position, a reversed trend emerged for T3 and T4, where high musicality participants produced higher F0_z values. For T3, this difference was most pronounced in the central dipping region of the pitch trajectory (see Fig. 49). For T4, a significant difference was also observed, with divergence between groups emerging from approximately Point 4 to Point 9.5 (see Fig. 50). No significant musicality effects were found for T1 or T2 in the second syllable.

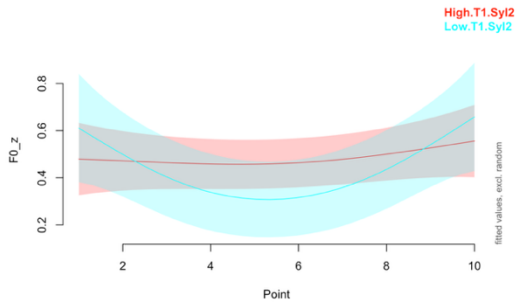


Figure 47 Tone 1 production on syllable 2 by Musicality

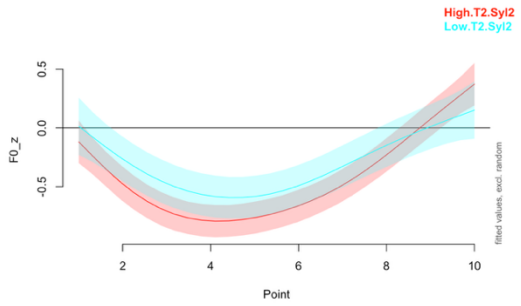


Figure 48 Tone 2 production on syllable 2 by Musicality

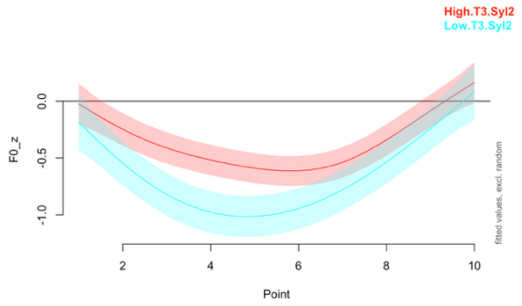


Figure 49 Tone 3 production on syllable 2 by Musicality

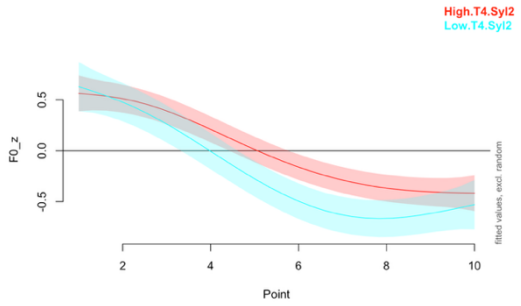


Figure 50 Tone 4 production on syllable 2 by Musicality

4.4.5 Interim summary on disyllabic targets

This section examined the production of Mandarin lexical tones in disyllabic target phrases in isolation by Italian learners of Mandarin, with a focus on how tone category, syllable position, and individual-level differences modulate pitch trajectories.

The initial analyses confirmed robust effects of both tone identity and syllable position on F0 realization. As expected, Mandarin tones displayed distinct pitch contours, mostly consistent with their canonical tonal shapes. T1 – the high-level tone – exhibited a stable F0 trajectory in syllable-initial position, but showed slightly higher variability in syllable-final contexts, likely due to phrase-final effects. In contrast, T2, T3, and T4 displayed more complex, time-varying contours, reflecting their phonological characteristics, respectively. Across all tones, tokens in syllable-final position displayed greater contour complexity yet converged

toward citation-like realizations, indicating reduced divergence from canonical tone shapes at phrase boundaries.

Among individual-level factors, Proficiency proved to be the strongest predictor of tonal F0 variation in disyllabic words production in isolation. Musical aptitude also enhanced predictive power, albeit to a lesser degree; whereas grade level (academic year) did not improve model fit, implying that formal academic progression may not reliably reflect gains in phonetic accuracy in disyllabic tonal production. This contrasts with findings from the monosyllabic analysis, where grade-level effects were more prominent.

Proficiency-related effects were particularly notable for T3 and T4. In syllable-initial position, high-proficiency learners produced significantly lower F0_z for T3, aligning more closely with its expected dipping contour. In syllable-final position, high-proficiency speakers produced higher F0_z values for both T3 and T4.

The influence of musicality followed a similarly tone- and position-specific pattern. For the first syllable, participants with higher musical sophistication produced significantly lower F0_z for T1 and T4; for T4, specifically the lowering in high proficiency learners lies on the second portion of the predicted curve. However, in the second syllable, the direction of the effect reversed: high-musicality participants produced significantly higher F0_z for T3 and T4, particularly in regions corresponding to the dipping and falling portions of the contour.

Interestingly, T2 did not reveal significant effects for either Proficiency or Musicality in either syllable position. Similarly, T1 and T3 did not show significant differences in the first syllable under the musicality model.

4.5 Discussion

Taken together, the results provide foundational insights into how Italian learners acquire and produce Mandarin lexical tones in isolation, and how this ability is modulated by individual-level factors such as language proficiency, musical aptitude, and academic progression. Within monosyllabic targets, despite the modest group and individual differences, learners largely produced monosyllabic targets with mostly appropriate pitch scaling and contouring, indicating that citation-form tonal categories are well established among participants' cohort. Yet, important asymmetries emerged. Tone 1 (high-level) was realized with the greatest consistency, while Tone 3 (dipping) and Tone 4 (falling) demonstrated the greatest variability across learners. Critically, while low-proficiency learners already approximate tonal shapes to a degree, higher proficiency was associated with finer-grained

control, especially for T4, where learners' increased pitch height in the early portion of the contour, suggesting improved planning and phonologically accurate execution of the falling trajectory. Musical aptitude further differentiated learners' performance: high-musicality individuals showed enhanced realization of dynamic tonal contours, particularly in the final portion of T4, reinforcing the view that auditory and sensorimotor pitch tracking skills support L2 tone acquisition.

The disyllabic results point to a tightly constrained, tone- and position-dependent system in which boundary conditions and individual differences shape learners' pitch implementation in predictable ways. Tone identity and syllable position exerted robust effects: T1 was stable in syllable-initial position but slightly more variable phrase-finally, whereas T2, T3 and T4 were characterized by richer time-varying dynamics; nonetheless, all tones tended to converge toward citation-like realizations at the phrase boundary, suggesting boundary-driven stabilization of tonal targets. Among learner variables, Proficiency emerged as the most informative predictor, outperforming Grade and – more modestly – Musicality. Proficiency effects clustered on the more dynamic tones: high-proficiency learners lowered T3 in the first syllable (a closer dip) and raised F0 in the second syllable for both T3 and T4. Musicality contributed in a similarly selective, position-dependent fashion: lower F0 for T1 and T4 in syllable-initial position, but higher F0 for T3 and T4 phrase-finally, particularly over the dip/fall regions. By contrast, T2 was comparatively stable, showing no reliable modulation by Proficiency or Musicality in either position. Taken together, these patterns suggest that learners' advancement in disyllabic target production in isolation depends less on their formal academic stage than on their ability to modulate tonal scaling and temporal alignment relative to syllable position. This aligns with the observation that phonological training in L2 Mandarin in Italy is often restricted to monosyllabic targets during the first year of study, with comparatively less attention devoted to disyllabic phrases, even in isolation (Francolino, 2022). The most substantial gains – and challenges – are observed for non-level tones in phrase-final contexts.

These findings have important implications for models of tone acquisition and for subsequent experimental work. While the current study was limited to isolated target words, it offers a controlled baseline against which tone production in prosodically rich environments can be interpreted. That learners exhibit substantial contour accuracy in isolation, yet struggle with modulation across syllable positions, invites the hypothesis that prosodic integration – not isolated tone production per se – is the key locus of difficulty in later stages of L2 tonal development. As Studies 2 and 3 will address disyllabic tone production in prosodic contexts, the patterns observed here suggest that learners may not only need to retrieve tonal categories

but also dynamically adjust them in response to prosodic phrasing, focus, and sentence type. The attenuated pitch movements in final position and the differential sensitivity of T3 and T4 to proficiency and musical aptitude anticipate potential challenges in deploying tones within complex intonational structures. Thus, these baseline findings position us to evaluate in subsequent analyses whether the tone-intonation interface constitutes a persistent bottleneck in L2 tonal prosody, and whether domain-general auditory skills can scaffold not only tone shape realization but also its functional deployment across linguistic contexts.

In sum, addressing RQ1-RQ2, the baseline models yielded evidence that Italian learners' F0-normalized trajectories in isolation largely track the canonical shapes of the four Mandarin tones: T1 is the most stable high-level contour, while T3 and T4 exhibit the richest time-varying dynamics and the greatest inter-speaker variability; T2 is comparatively stable across speakers and contexts. In monosyllables, learners generally achieve appropriate pitch scaling and contouring, indicating well-formed citation categories – supporting H1. In disyllables, tone and syllable position exert robust, systematic effects: T1 remains steady in initial position, and all tones display more complex dynamics phrase-finally; nevertheless, contours tend to converge toward citation-like realizations at the boundary, consistent with a boundary-driven stabilization – partially supporting H3. Accuracy is therefore not uniformly lower in disyllables, but rather tone- and position-dependent, with increased demands for T3 and T4 – broadly in line with H2. Individual differences sharpen these patterns: Proficiency is the strongest predictor (e.g., better dip execution in T3 and higher early F0 for T4 among high-proficiency speakers), while musicality selectively benefits dynamic portions (notably the final descent of T4). Together, these results provide the required baseline against which prosodically modulated disyllables can be compared, and they motivate the next studies' focus on how learners adjust tonal scaling and timing under phrase-level constraints (e.g., focus, sentence type) beyond isolated production.

5 Prosodic Encoding of Focus in L2 Mandarin: how tone and focus interact

This study presents a systematic investigation of the interactive prosodic encoding of lexical tone and contrastive focus in disyllabic L2 Mandarin productions, with particular attention to their realization by Italian learners. Through a controlled reading task, we compare native speakers (n = 10) and Italian learners (n = 42) in their pitch-based strategies for focus marking, using acoustic measures derived from fundamental frequency (F0) – the primary acoustic correlate of prosodic prominence and tonal identity in Mandarin (Xu, 1999; Chen, 2010).

5.1 Research questions and hypotheses

The overarching goal of this study is to examine whether, and in what ways, Italian university learners of Mandarin encode focus in ways that diverge from native-speaker norms, and to determine whether such divergences may stem from cross-linguistic transfer from Italian. The findings are expected to yield insights of both theoretical and pedagogical relevance, particularly regarding prosodic instruction in L2 Mandarin. More specifically, this study addresses the following research questions (RQs):

RQ1. How does focus condition (on-focus vs. pre-/post-focus) modulate F0 scaling and contour shape as a function of lexical tone (T1-T4) and syllable position within disyllabic targets for L1 vs. L2 speakers?

RQ2. Do Italian learners implement canonical Mandarin focus mechanisms (e.g., on-focus enhancement and post-focus compression) to a native-like degree across tones and positions? Are these differences systematically predicted by Grade, Proficiency, and Musicality?

RQ3. To what extent do Italian learners' focus strategies exhibit intonational transfer from their L1, e.g. through the use of falling contours as a default marker of emphasis, regardless of the lexical tone involved?

The experimental hyphothesis (H) are as follows:

H1. Native speakers are expected to exhibit tone-specific patterns of focus realization, characterized by elevated or expanded F0 on the focused syllable and robust post-focal compression on subsequent material, accompanied by contour adjustments appropriate to each tone (e.g., an enhanced rise for T2, a steeper fall for T4, and controlled scaling for T1 and T3; Xu, 1999; Chen, 2010). These effects are predicted to surface even within the minimal intonational phrases used in the experimental materials.

H2. Italian learners are expected to exhibit attenuated and less tone-specific realizations of focus, relying predominantly on global F0 height rather than tone-appropriate contour modulation. As a result, their productions are predicted to exhibit weaker post-focus compression and reduced differentiation across focus conditions.

H3. Learners will display L1-driven strategies incongruent with Mandarin norms, such as a generalized falling tendency to signal emphasis (irrespective of lexical tone), consistent with salient emphasis cues in Italian intonation (cfr. Sbranna et al., 2023; *inter alia*).

H4. Focus effects are expected to interact with syllable position: in native speakers, phrase-final syllables should exhibit stronger boundary-related modulation, whereas learners are predicted to only partially implement these position-sensitive adjustments.

Together, these predictions test whether L2 speakers not only reach appropriate tonal targets but also integrate tone with information-structure cues in a manner that is both tone-specific and context-sensitive.

5.2 Dataset Overview

The dataset includes annotated syllables extracted from the main experimental task (see § 3.3) and includes both L1 Mandarin and Italian L2 Mandarin speakers. As already mentioned, only statements were included for this analysis. Each syllable was uniquely identified and annotated for linguistic and prosodic variables as follows:

- SylID: Unique identifier for each syllable;
- Speaker: Individual speaker identifier;
- Lang: Language group (CH = native Mandarin speakers; IT = L2 Mandarin learners);
- SylPos: Position of the syllable in the disyllabic phrase (Syl1, Syl2);
- Tone: Lexical tone of the target syllable (T1-T4);
- OtherTone: The lexical tone of the adjacent syllable (preceding or following, depending on position);
- Focus: Focus condition (on-focus, pre-focus, post-focus);
- F0: Raw F0 values (in Hz) extracted at 10 time-normalized points per syllable;
- F0_z: z-score normalized F0 values per speaker.

For each syllable, a set of derived F0 curve parameters was calculated, intended to capture the shape, level, and dynamism of the pitch contour. These measures are well-established in

prosodic research as indicators of both lexical tone realization and focus-induced prominence (Xu, 2013; Linke et al., 2020):

Table 26 F0 Curve parameters analyzed in the study

Measure	Description
Mean F0	Mean F0_z across the 10 sampled points
F0_max	Maximum F0_z value
F0_min	Minimum F0_z value
F0_range	Difference between F0_max and F0_min
F0_slope	Linear slope coefficient from F0_z ~ time (Point) model

To capture potential differences in tonal and focus realization between native and non-native speakers, the data were analyzed separately for each syllable position (Syl1 and Syl2). This decision was motivated both by the findings of Study 1 (see § 4) and by previous research showing that prosodic cues – particularly those related to focus marking and tonal implementation – vary systematically as a function of syllable position in Mandarin (Chen, 2006; Hsu & German, 2018; Athanasopoulou et al., 2019). Accordingly, two parallel analyses were conducted, one for each syllable position. For qualitative reference, Appendix H presents by-speaker contour plots for both the Syl1 and Syl2 datasets.

5.3 First-syllable analysis

To investigate how pitch contours in monosyllabic productions are shaped by the interplay of language background, lexical tone, and focus condition, a GAMM was fitted to the normalized pitch data (F0_z) using the *bam()* function from the *mgcv* package. The model included a full three-way interaction between Language (native Mandarin speakers vs. Italian learners), Tone (T1-T4), and Focus (on-focus vs. pre-focus), with smooth terms estimated over normalized syllable time (Point) for each Language.Tone.Focus combination. To control for speaker-specific variability and potential effects from adjacent tones (i.e., tonal coarticulation), random smooth terms were added for Speaker and OtherTone, respectively. This structure allowed the model to flexibly capture complex pitch contours and contour modulations across experimental conditions[17].

[17] The model successfully converged under the fREML criterion and demonstrated a good overall fit, explaining 31.5% of the deviance (adjusted R2=0.307) with a scale parameter estimate of 0.55. Diagnostic evaluations confirmed that the selected basis dimensions (k=10) were sufficient for capturing the observed contour complexity,

To assess how language background and focus condition modulate pitch within each lexical tone, post hoc pairwise comparisons were conducted using EMMs. Below, relevant contrasts for each target tone are reported.

5.3.1 Tone 1

The analysis revealed a significant effect of focus condition on F0_z for native Mandarin speakers, who produced higher F0_z values in the on-focus condition relative to pre-focus syllables. Italian learners demonstrated a similar pattern of prosodic modulation; however, the magnitude of the contrast between on-focus and pre-focus F0_z values was comparatively reduced. No statistically significant group differences were observed in either the on-focus or pre-focus conditions, indicating that both native and non-native speakers modulated pitch height in response to focus.

A summary of key pairwise contrasts for T1 in Syl1, comparing F0_z values across focus conditions and language groups, is provided in Tab. 27:

Table 27 F0_z pairwise contrasts for Tone 1 in syllable 1

Contrast	Estimate	SE	*t*	*p*
CH.on - CH.pre	+0.194	0.0595	3.26	0.0061
IT.on - IT.pre	+0.088	0.0302	2.93	0.0181
CH.on - IT.on	+0.032	0.096	0.33	n.s.
CH.pre - IT.pre	-0.074	0.093	-0.79	n.s.

These findings suggest that both groups are sensitive to the prosodic encoding of focus in Mandarin, at least in terms of F0 register on the first syllable bearing T1. Nonetheless, native speakers exhibited a more pronounced pre-focus lowering effect. This pattern is consistent with previous descriptions of Mandarin focus, in which F0 suppression in pre-focus regions contributes to enhancing perceptual salience at the focus site (see § 2.6.2).

Predicted GAMM curves reveal a more stable and flatter F0 contour across both focus conditions in the productions of native Mandarin speakers. Particularly noteworthy is the final portion of the on-focus contour: native speakers maintain a high F0 throughout the syllable, with minimal decline toward the offset. In contrast, the productions of Italian learners exhibit

with no evidence of undersmoothing (all k-indices ≈ 0.98; all associated p-values > 0.035). This suggests that the model's smoothing structure adequately represented the shape and variation of the F0 trajectories across conditions.

a more pronounced decline in F0 across the syllable, with noticeably lower F0 values at the final portion relative to the initial portion. This pattern results in a more falling-like contour, diverging from the smoother, sustained pitch trajectory observed in native productions.

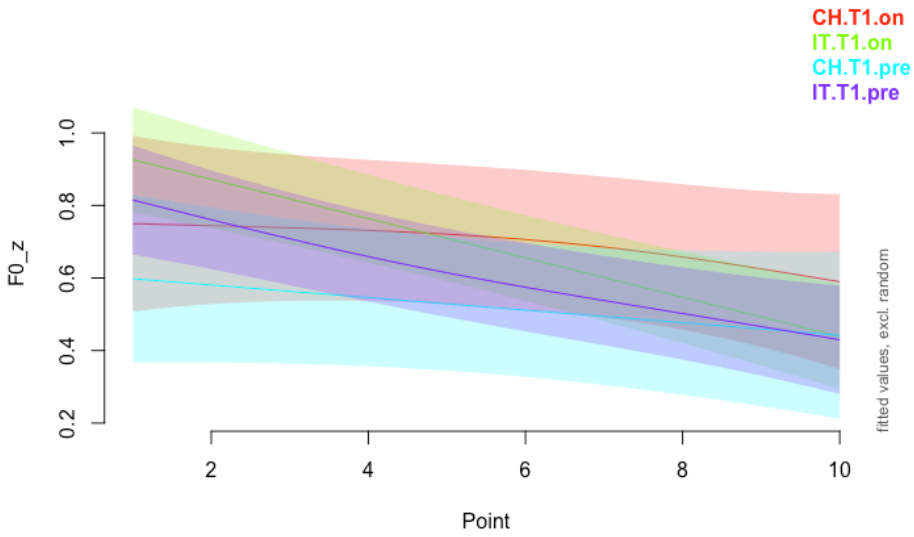


Figure 51 Tone 1 production by Language and Focus

5.3.2 Tone 2

The analysis of T2 revealed robust and statistically significant effects of both Language and Focus condition on pitch realization, pointing to notable divergences in how L1 and L2 speakers encode prosodic prominence.

Parametric comparisons of F0_z for T2 revealed a robust effect of both focus condition and L1 group (see Tab. 28). In particular, native Mandarin speakers (CH) produced significantly lower F0_z values in on-focus contexts compared to both their own pre-focus productions and those of Italian learners (IT). Specifically, the contrast between CH.on and IT.on was highly significant, as was CH.on vs. CH.pre, indicating that focus was associated with a reduction in F0_z height among native speakers. Italian learners, by contrast, showed significantly higher pitch values in on-focus conditions, suggesting a more straightforward F0-raising strategy.

Table 28 F0_z pairwise contrasts for Tone 2 in syllable 1

Contrast	Estimate	SE	df	*t*-ratio	*p*-value	Significance
CH.on - CH.pre	-0.3017	0.0692	16444	-4.358	0.0001	***

Contrast	Estimate	SE	df	*t*-ratio	*p*-value	Significance
IT.on - IT.pre	+0.1390	0.0456	16444	3.052	0.0122	*
CH.on - IT.on	-0.5749	0.1060	16444	-5.439	< .0001	***
CH.pre - IT.pre	-0.1342	0.0958	16444	-1.401	0.4985	n.s.

These results are further corroborated by the GAMM-derived F0_z trajectories. As visualized in Fig. 52, native speakers' on-focus T2 contours (red curve) consistently lie below all other curves across the majority of the syllable, confirming the lowering of F0_z in CH.on contexts. Italian learners' on-focus realizations (green curve), by contrast, are visibly higher in pitch throughout.

This pattern may seem counterintuitive under traditional models that associate focus with F0 elevation; however, it aligns with previous findings indicating that Mandarin T3 – and potentially T2 – can be lowered under focus conditions (cf. Xu, 1999). One possible explanation is that native speakers prioritize contour enhancement over pitch height, lowering the onset to maximize the perceptual salience of the rising movement in T2. Italian learners, lacking this level of tonal-prosodic integration, instead apply a global F0 boost that flattens the contour and potentially reduces perceptual distinctiveness.

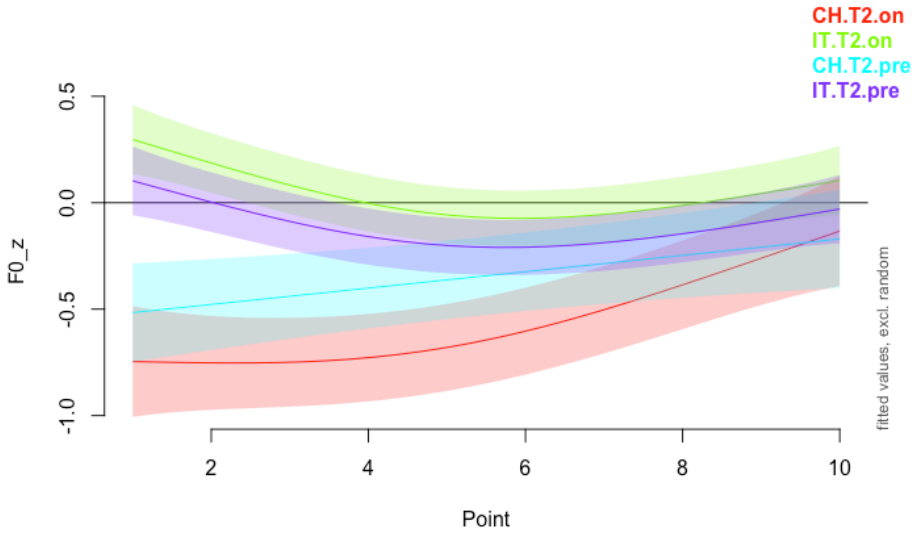


Figure 52 Tone 2 production by Language and Focus

5.3.3 Tone 3

The analysis of T3 revealed a pattern that closely parallels the findings for T2, highlighting a consistent divergence in pitch realization between native speakers and L2 learners. In both the on-focus and pre-focus conditions, Italian learners produced markedly higher F0_z values than their native Chinese counterparts, suggesting a systematic difficulty in targeting the low F0 register characteristic of T3 (see Tab. 29).

Specifically, the on-focus realizations of T3 by Italian learners exhibited significantly elevated F0 compared to native productions. A similar pattern was observed for pre-focus syllables, where L2 speakers again displayed higher F0_z values than native speakers. These differences suggest that Italian learners failed to reach the low F0 targets required for T3 in both prosodic contexts, an issue that may reflect L1 transfer from intonation languages like Italian, where low pitch excursions are comparatively less frequent and less functionally significant on lexical level.

Table 29 Key pairwise contrasts by Language

Contrast	Estimate	SE	*t*	*p*
CH.on - IT.on	-0.390	0.0952	-4.10	***0.0002
CH.pre - IT.pre	-0.395	0.100	-3.95	***0.0005

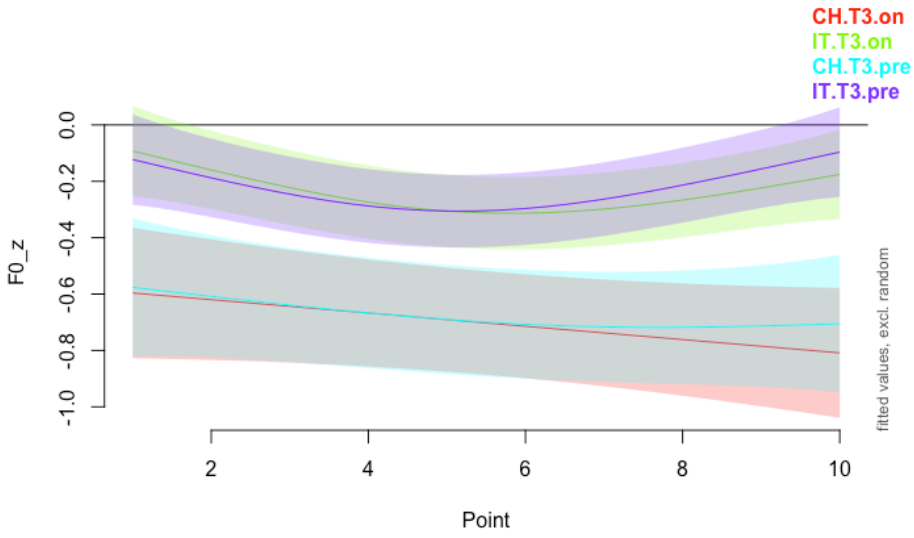


Figure 53 Tone 3 production by Language and Focus

Unlike the more complex interactions observed for T2, T3 did not show significant modulation by focus within each language group in this experimental conditions. Instead, the

dominant source of variation appears to be the overall pitch scaling across speaker groups, rather than within-group prosodic contrast.

5.3.4 Tone 4

The analysis of T4 revealed a more nuanced pattern of results compared to the other tones. While there was a significant language effect in the on-focus condition – with CH producing higher F0_z than IT (*estimate* = +0.276, *SE* = 0.104, *t* = 2.64, *p* = 0.041) – other pairwise contrasts did not reach significance.

These data suggest that T4 in Syll is subject to more modest modulation under focus, particularly among L2 speakers. Visual inspection of the GAMM smooths supports this interpretation: Italian learners produced less steep falling contours, especially in on-focus positions, which may indicate a reduction in tonal dynamicity associated with this tone.

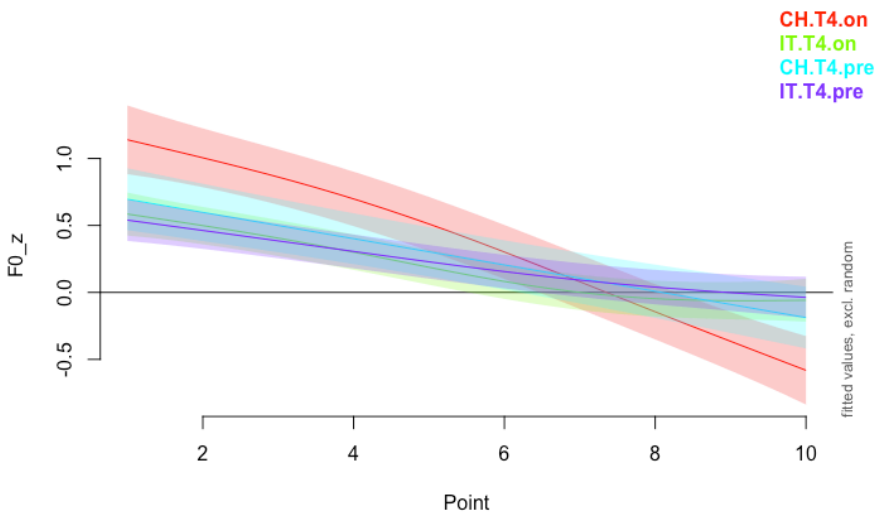


Figure 54 Tone 4 production by Language and Focus

Notably, the absence of significant pre- vs. on-focus differences among Italian learners again points to a lack of prosodic differentiation. While native speakers exhibit consistent pitch lowering in pre-focus contexts, Italian L2 speakers do not appear to consistently implement this contrast (see also F0_slope and F0_range results). The modulation of T4 in L2 productions under focus conditions on Syll, therefore, seems to be limited both in scope and in gradient realization.

5.3.5 Analysis on Curve Parameters for Syl1

5.3.5.1 Mean F0

To further assess potential differences in pitch height during the production of the first syllable (Syl1), we analyzed the mean fundamental frequency (Mean F0), z-score normalized within speaker, across all tokens. A GLMM was fitted with Language (Mandarin L1 vs. Italian L2), Lexical Tone (T1-T4), and Focus condition (on-focus vs. pre-focus) as fixed effects. Random intercepts were included for Speaker and OtherTone to account for individual variation and tonal coarticulatory context, respectively.

Model comparison using likelihood ratio tests confirmed that both random effects significantly improved model fit, thereby supporting their inclusion in the final model structure:

Table 30 Random effects significance through model comparison

Random Effect	**ΔAIC**	**LRT**	***p***
Speaker	+164.2	166.13	*** <.001
OtherTone	+2.9	4.83	* =.028

Initial inspection of the fixed effects revealed main effects of Language and Tone, as well as a significant Language.Tone interaction. However, neither the main effect of Focus nor the three-way interaction (Language.Tone.Focus) reached significance. After model simplification via backward selection, the final model retained Language, Tone, and Focus as main effects, along with the Language.Tone interaction.

The effects of Tone and Language are visualized in Fig. 55, which plots the EMMs of mean F0 by tone and language group.

Importantly, a clear divergence emerged for T2 and T3 across Language groups: IT produced significantly higher mean F0 values for these tones than CH, suggesting potential L2-related difficulties in accurately implementing target pitch contours. In contrast, T4 revealed a smaller language effect.

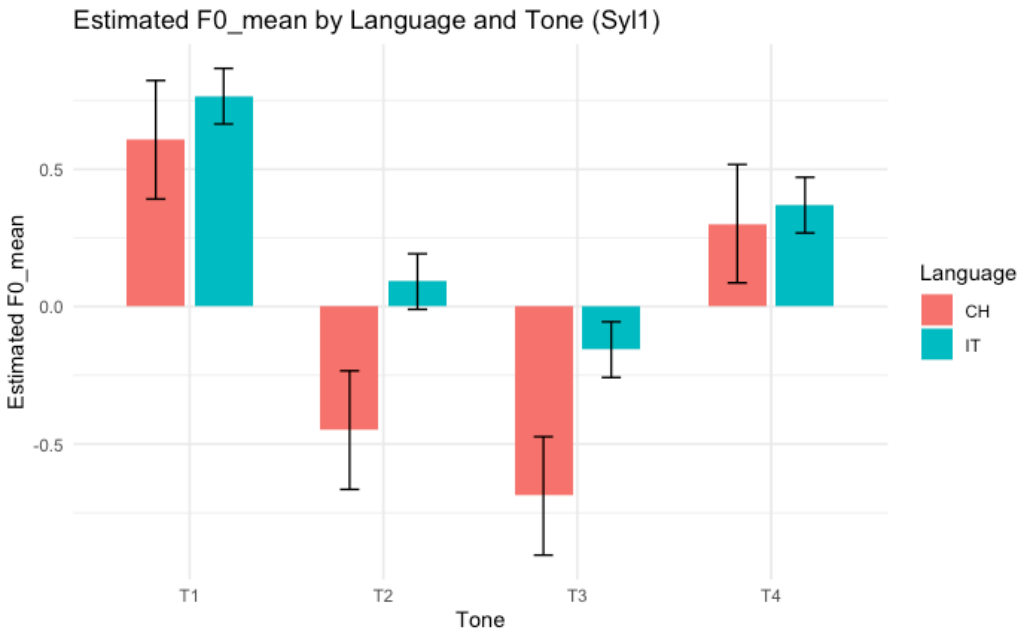


Figure 55 Estimated Mean F0 by Language and Tone (Syl1)

The effect of focus on pitch height was relatively modest but aligned with prosodic expectations. As illustrated in Fig. 56, syllables produced in on-focus contexts consistently exhibited slightly elevated mean F0 compared to their pre-focus counterparts, across both CH and IT, and across all lexical tones. This pattern is broadly consistent with prosodic theories of focus marking, wherein pitch elevation serves as a primary acoustic correlate of prominence (Xu, 1999; Chen & Gussenhoven, 2008). While the magnitude of this elevation varied across tones, the directionality of the effect supports findings from prior studies in Mandarin and other tonal languages, which report increased F0 and expanded pitch range in focused constituents (see § 2.6.2). These results suggest that even L2 learners, despite ongoing difficulties with tonal contrast, are sensitive to the prosodic demands of focus marking and attempt to encode pragmatic meaning through pitch modulation.

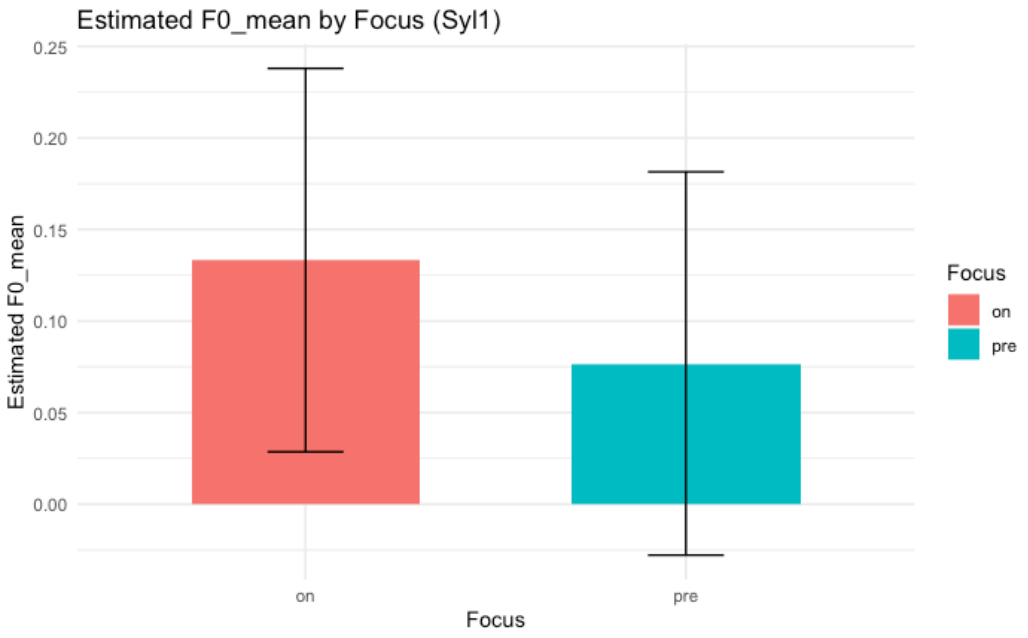


Figure 56 Estimated Mean F0 by Focus (Syl1)

In sum, this analysis confirms that mean F0 is systematically shaped by lexical tone, with clear tonal distinctions across all speakers on Syl1. Additionally, Language background exerts a significant effect, particularly for T2 and T3, where IT produced elevated pitch levels relative to CH. Although the effect of focus on mean pitch was relatively small, it followed the expected direction, at least for T1 and T4, with slightly higher F0 values in on-focus conditions.

These findings suggest that while tonal realization is broadly preserved in L2 production, certain tones may pose greater challenges, and focus marking via pitch may be incompletely or subtly encoded by L2 learners.

5.3.5.2 F0 slope

To investigate pitch movement across the first syllable of the disyllabic target phrases, we examined the F0 slope, a parameter representing the direction and steepness of pitch change over the syllable duration. This measure is particularly informative for tones with inherently rising or falling contours, such as Mandarin T2 and T4, and is also sensitive to prosodic prominence effects introduced by focus.

A GLMM was fitted with Language (CH vs. IT), Tone (T1-T4), and Focus (on-focus vs. pre-focus) as fixed effects, along with all possible interactions. Speaker and OtherTone were included as random intercepts to account for inter-speaker variability and tonal coarticulation effects stemming from adjacent syllables.

Likelihood-ratio tests demonstrated that both random effects significantly improved model fit, justifying their inclusion in the final structure:

Table 31 Random effects significance through model comparison

Random Effect	ΔAIC	LRT	*p*
Speaker	+44.9	46.96	*** < .001
OtherTone	+12.2	14.25	*** < .001

Following model simplification, the final model retained the main effects of Tone, Tone.Focus, Language.Tone, and importantly, a three-way interaction between Language, Tone, and Focus.

As expected, T2 was associated with a positive slope, reflecting its canonical rising pitch contour, while T4 showed a strongly negative slope, consistent with its falling contour.

However, the presence of significant interactions reveals more nuanced variation across groups. The Language.Tone interaction indicates that slope values differed between CH and IT for certain tones. Additionally, the Tone.Focus interaction confirms that pitch slope is sensitive to prosodic structure, particularly whether a syllable is on focus or not.

Crucially, the three-way interaction between Language, Tone, and Focus underscores that these two domains – tonal identity and discourse status – are not realized independently, and that non-native speakers do not modulate pitch slope in a fully native-like manner: the interaction was primarily driven by T4, where IT failed to maintain the steep falling slope in pre-focus conditions.

Specifically, the interaction between language background, tone category, and focus placement yielded a statistically significant effect ($\beta = -0.102$, $p = .009$), indicating that IT produced significantly flatter F0 contours for T4 in pre-focus contexts compared to CH. This pattern suggests that L2 learners may struggle to maintain the prosodic distinctiveness of lexical tones when simultaneously encoding information-structural cues such as focus. Unlike native speakers, who preserved the canonical sharp falling contour of T4 even in pre-focus positions, L2 learners exhibited a compression or flattening of the pitch slope on Syl1. One possible interpretation is that this flattening reflects a form of prosodic economy, wherein learners suppress tonal modulation on the pre-focused syllable in anticipation of increased pitch dynamism on the upcoming focused constituent. This preparatory adjustment may reflect a limited capacity to manage competing prosodic demands.

The interaction was further examined through EMMs, plotted across focus conditions for each tone and language group. As reported in Fig. 57, native speakers preserved the expected falling contour of T4 in pre-focus contexts, while Italian learners exhibited a marked reduction in slope steepness. This contrast is absent or less pronounced in the other tones, especially T2, where both groups produced similar rising slopes.

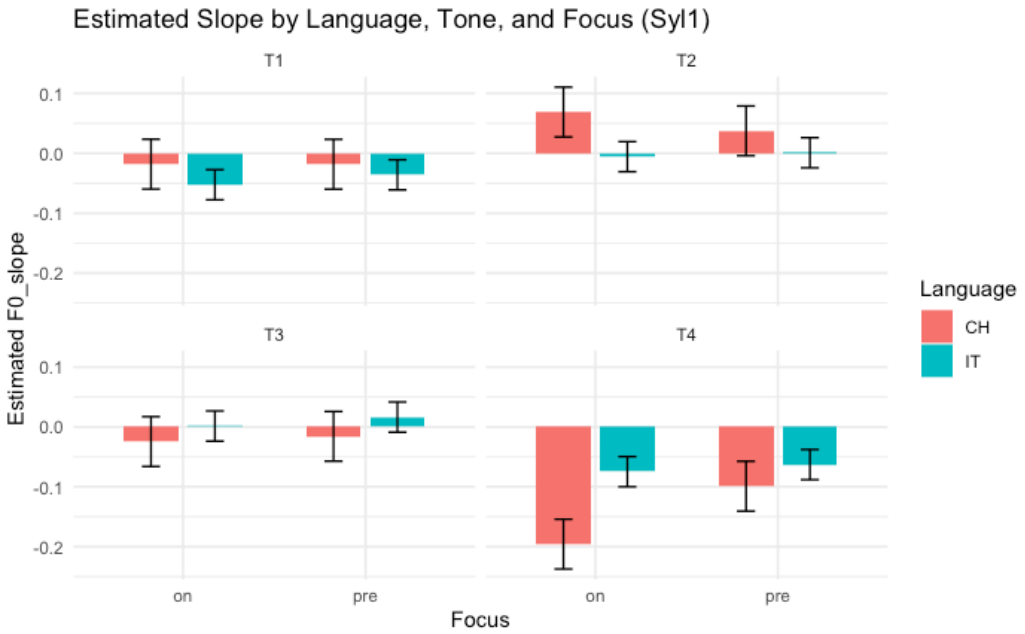


Figure 57 Estimated Slope by Language, Tone, and Focus (Syl1)

These findings highlight the phonetic complexity of slope realization in L2 tone production. While both native and non-native speakers produced rising slopes for T2, the Italian learners exhibited reduced pitch modulation in T4, especially under pre-focus conditions. This failure to dynamically adjust pitch according to focus suggests a limited prosodic repertoire for managing interactions between tonal identity and discourse prominence.

The observed three-way interaction confirms that L2 speakers' pitch dynamics are shaped not only by their representation of lexical tone, but also by how they coordinate tonal and prosodic cues. Importantly, such coordination appears language-specific: while native speakers modulate slope flexibly across discourse contexts, L2 learners may adopt more flattened or compressed patterns, possibly reflecting constraints in their L2 phonological encoding.

These results underscore that pitch slope is a key parameter for understanding non-native tone realization within prosodic context, and that slope flattening – especially in contextually deaccented positions – may serve as a diagnostic feature of L2 prosodic competence.

5.3.5.3 F0 max

To assess how speakers encode prosodic prominence and tonal identity through pitch maxima, we analyzed the parameter F0_max – the highest point in the F0 contour – within the first syllable of the disyllabic target phrases.

A GLMM was fitted, predicting F0_max as a function of Language (L1 Mandarin vs. L2 Italian), Tone (T1-T4), and Focus condition (on-focus vs. pre-focus), including all two-way and three-way interactions among these fixed effects. Random intercepts were included for Speaker and OtherTone, to control for individual differences and possible coarticulatory effects due to tonal context.

Model comparison based on likelihood-ratio tests revealed that Speaker contributed significantly to model fit ($\chi^2 = 143.64$, $p < .001$), while OtherTone did not ($\chi^2 = 0.96$, $p = .33$), and was therefore excluded from the final model.

Subsequent model reduction via backward elimination led to the retention of the main effects of Language, Tone, and Focus, as well as a significant Language.Tone interaction. The three-way interaction between Language, Tone, and Focus, along with all other higher-order terms, did not significantly improve model fit, suggesting that focus-related increases in F0_max are expressed similarly across both native and non-native groups.

Tone exerted a strong main effect on F0_max, with T1 yielding the highest maxima, and T3 the lowest. This ordering aligns with the expected pitch targets of Mandarin lexical tones, thereby validating the F0_max parameter as a faithful phonetic reflection of tonal identity.

Importantly, the Language.Tone interaction ($p < .001$) indicates that IT diverged from CH in the way they scaled pitch maxima across tones. Of particular interest is the significant negative coefficient associated with the LangIT.ToneT4 interaction ($\beta = -0.392$, $p = .0029$), which suggests that IT produced lower F0 maxima for T4 than CH. This pattern may reflect a flattening or compression of the falling contour, potentially resulting in diminished distinctiveness for this tone.

Although interactions involving T2 and T3 were not statistically significant, there was a trend toward slightly elevated F0_max values in the L2 group, possibly indicating an overcompensation strategy when producing rising or dipping tones. These tendencies are visualized in Fig. 58, which displays EMMs of F0_max by Language and Tone. The plot illustrates a clear downward shift for T4 in the IT group, whereas CH maintains a wider pitch range, preserving distinct tonal scaling.

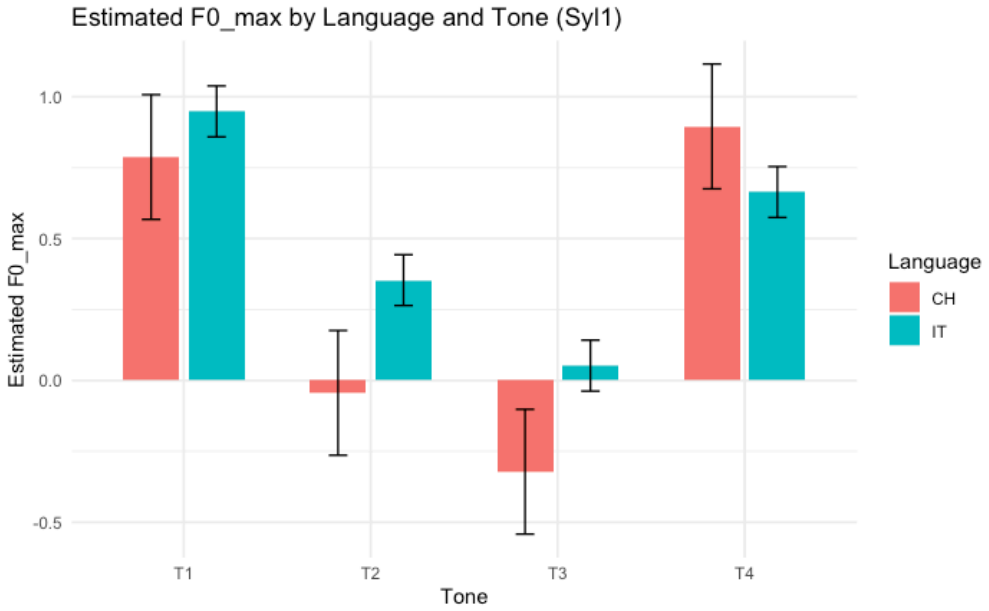


Figure 58 Estimated F0_max by Language and Tone (Syl1)

Across all tones and both speaker groups, Focus condition exerted a significant effect on F0_max ($p = .001$), such that focused syllables were produced with higher maximum pitch than pre-focused ones. These findings identify F0 peaks as a robust correlate of prosodic prominence in Mandarin. However, notably, no significant interaction emerged between Language and Focus, suggesting that IT modulate F0_max in focus marking similarly to CH. In other words, while tonal scaling patterns differ between the groups – particularly in T4 – the use of pitch maxima to mark focus is consistent across L1 and L2 production.

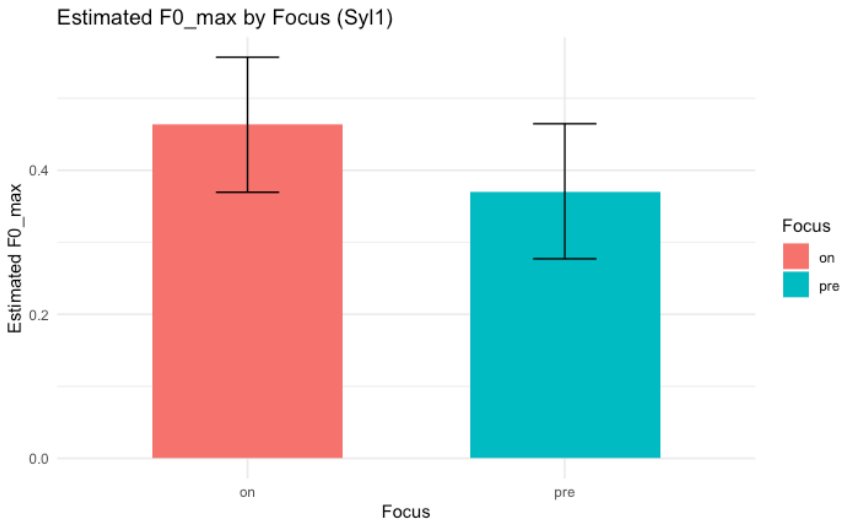


Figure 59 Estimated F0_max by Focus (Syl1)

Overall, the findings from the F0_max analysis reveal a nuanced picture. On one hand, maximum pitch is clearly sensitive to both tone category and prosodic focus, with native speakers exhibiting expected scaling patterns across tones and prominence levels. L2 speakers, by contrast, demonstrate less precise tonal scaling, particularly in T4, where F0 maxima are markedly reduced, possibly indicating limited control over falling contours.

Yet crucially, both groups appear to reliably use F0_max as a cue to prosodic focus, with increased pitch height in focused syllables. This suggests that while F0_max is a robust marker of focus across speaker groups, it may be insufficient as a diagnostic for L2 tonal accuracy, due to its relatively intact status in learners' prosodic systems. The more informative differences may lie in how F0_max is distributed across tones, rather than how it is used to encode focus.

5.3.5.4 F0 min

To explore whether minimum pitch (F0_min) contributes to the realization of prosodic focus, and whether this contribution differs between CH and IT, a GLMM was fitted to the F0_min values extracted from the first syllable of disyllabic utterances. The model included Language, Tone, and Focus as fixed effects, along with all interactions, and incorporated a random intercept for Speaker and OtherTone to account for individual variability. Likelihood-ratio tests confirmed the significance of the Speaker term ($\chi^2 = 117.27$, $p < .001$), while the random intercept for OtherTone did not improve model fit and was therefore excluded.

Following backward model selection, the optimal fixed-effects structure retained main effects of Language, Tone, and Focus, as well as significant two-way interactions between Language.Tone and Language.Focus. The three-way interaction (Language.Tone.Focus), however, was not statistically significant and was removed from the final model. The absence of this higher-order interaction suggests that the relationship between Focus and F0_min is not modulated by tone in a language-specific manner, but rather that language background and prosodic structure exert more general, independent effects.

The final model revealed several noteworthy patterns. First, a robust main effect of Language confirmed that IT consistently produced higher F0_min values across conditions compared to native speakers. This finding is indicative of a shallower pitch floor in L2 productions, which likely contributes to a compressed pitch range – a phenomenon previously observed in non-native tonal production.

Second, a main effect of Tone was observed, reflecting tonal distinctions in pitch floor. Unsurprisingly, T3, characterized by its low dipping contour, had the lowest F0_min value, while T4 showed a higher minimum pitch overall. This tonal ranking mirrors native pitch targets, though L2 speakers' performance deviated, particularly in the dynamic tones.

The main effect of Focus, while only marginal (p = .072), pointed to a trend of pitch floor lowering under focus: F0_min was generally lower in focused syllables than in pre-focused ones. This trend aligns with prior findings in Mandarin that suggest that prosodic focus enhances pitch range, in part by depressing the F0 minimum to increase perceptual salience.

Critically, however, the Focus effect was not uniform across speaker groups. The significant Language.Focus interaction (p = .030) reveals that CH modulate F0_min more strongly in response to focus than IT. While native Mandarin speakers clearly lowered their pitch floor under focus – expanding the pitch range and marking prominence – Italian speakers exhibited more constrained modulation, with smaller differences in F0_min between focus conditions. This suggests a reduced use of pitch floor lowering as a cue to prosodic focus in L2 productions.

Moreover, the Language.Tone interaction confirmed that L2 speakers' elevated F0_min values were not uniform across tones, but were particularly pronounced for T2 and T3 – tones that involve rising and dipping contours and thus demand more pitch movement. These findings point to an overall reduction in pitch dynamism in L2 productions, particularly in tones that require large F0 excursions.

These differences are illustrated in Fig. 60, which plots EMMs of F0_min across Focus conditions for each tone and speaker group. The figure demonstrates that CH consistently lower F0_min under focus, while IT maintain higher minima and show limited focus-related modulation.

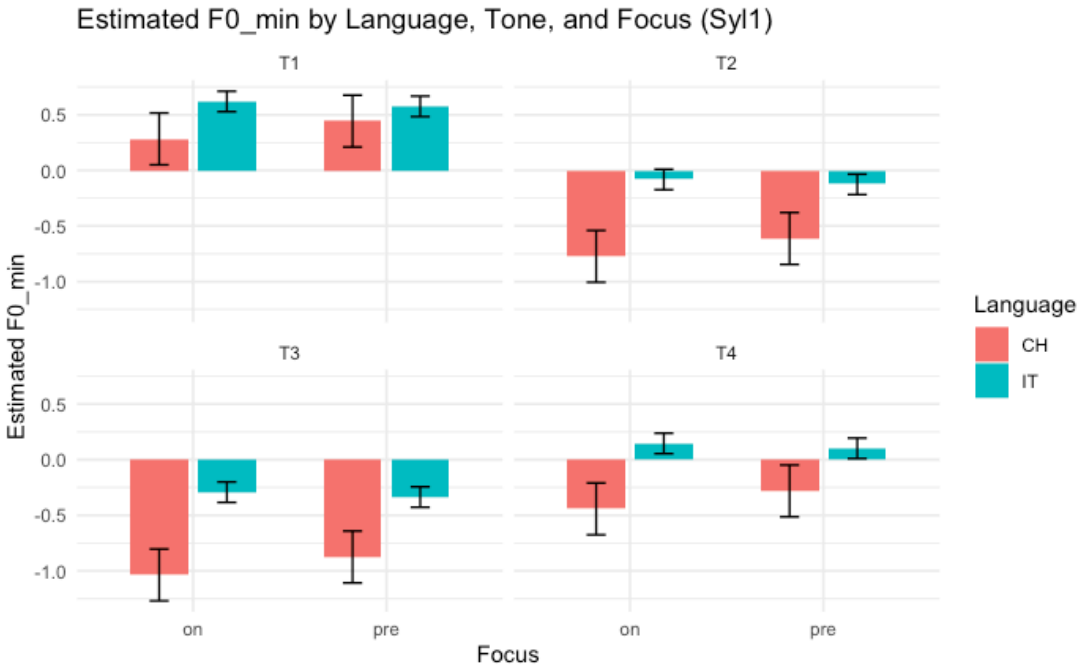


Figure 60 Estimated F0_min by Language, Tone, and Focus (Syl1)

Taken together, the results from the F0_min analysis support the interpretation that minimum pitch serves as a meaningful correlate of prosodic focus in native Mandarin productions on Syl1. The lowering of the pitch floor under focus conditions reflects a strategy of pitch range expansion, enhancing tonal contrast and prominence. In contrast, L2 Italian learners appear to adopt a more conservative pitch strategy, characterized by elevated minima and reduced modulation, particularly in tones that rely on dynamic pitch movement.

Although F0_min alone may not fully capture the complexity of L2 prosodic marking, its behavior reveals a clear difference in how L1 and L2 speakers exploit the lower end of the pitch spectrum. This diminished use of pitch floor manipulation may contribute to weaker perceptual marking of focus and less robust tonal realization in L2 Mandarin speech.

5.3.5.5 F0 range

This section investigates whether F0_range – calculated as the difference between F0_max and F0_min within the first syllable – serves as a robust prosodic cue for focus marking, and whether this acoustic parameter is deployed differently by L1 Mandarin speakers and L2 learners.

A GLMM was constructed to predict F0_range as a function of Language (CH vs. IT), Lexical Tone (T1-T4), and Focus condition (on-focus vs. pre-focus), along with their interactions. The model included a random intercept for Speaker, whose inclusion significantly

improved model fit ($\chi^2 = 11.77$, $p < .001$). By contrast, the random effect of OtherTone did not contribute significantly and was excluded from the final model.

Following backward stepwise model selection, the final model retained the full three-way interaction among Language, Tone, and Focus, confirming that the interplay of these three factors meaningfully shapes pitch range modulation ($F(3, 2945.7) = 3.21$, $p = .022$).

The results reveal a consistent pattern: F0_range increases under focus, confirming its role as a core prosodic cue. The main effect of Focus was significant ($p < .001$), indicating that, across speakers and tones, focused syllables are produced with wider pitch excursions – a finding in line with previous research on Mandarin prosody (see § 2.6.2).

However, the interpretation of this focus-driven range expansion is nuanced. While Focus was significant as a main effect, the non-significant estimate for Focus.pre in post-hoc contrasts suggests that the effect is modulated by interactions with Tone and Language, rather than being uniform across conditions.

Surprisingly, the main effect of Language was not significant at the coefficient level ($p = .590$), suggesting that global F0_range does not differ dramatically between the two speaker groups. Nonetheless, this should not obscure the more telling pattern of significant interactions – especially Language.Focus ($p = .001$), Tone.Focus ($p = .0018$), and the crucial three-way interaction ($p = .022$). These interactions reveal that L2 speakers do not employ pitch range for focus marking in the same way as L1 speakers, particularly within specific tonal contexts.

Tone identity had a strong influence on F0_range, as expected. T4 exhibited the widest pitch range, consistent with its phonological specification. However, the extent to which this range expanded under focus differed significantly by speaker group. The Language.Tone.Focus interaction revealed that L2 speakers showed markedly attenuated range expansion in T4, as reflected in the significant interaction terms: LangIT.ToneT4 ($p < .001$) and LangIT.ToneT4.Focuspre ($p = .003$).

These findings suggest that L2 speakers do not fully exploit the prosodic space available to them in tonal contexts, particularly where tonal cues require substantial pitch modulation, such as in T4. The failure to increase pitch range under focus in these tones points to limited pitch control or incomplete mastery of the tone-intonation interface. This pattern is clearly illustrated in Fig. 61, which plots EMMs of F0_range across Focus conditions, separately for each tone and language group.

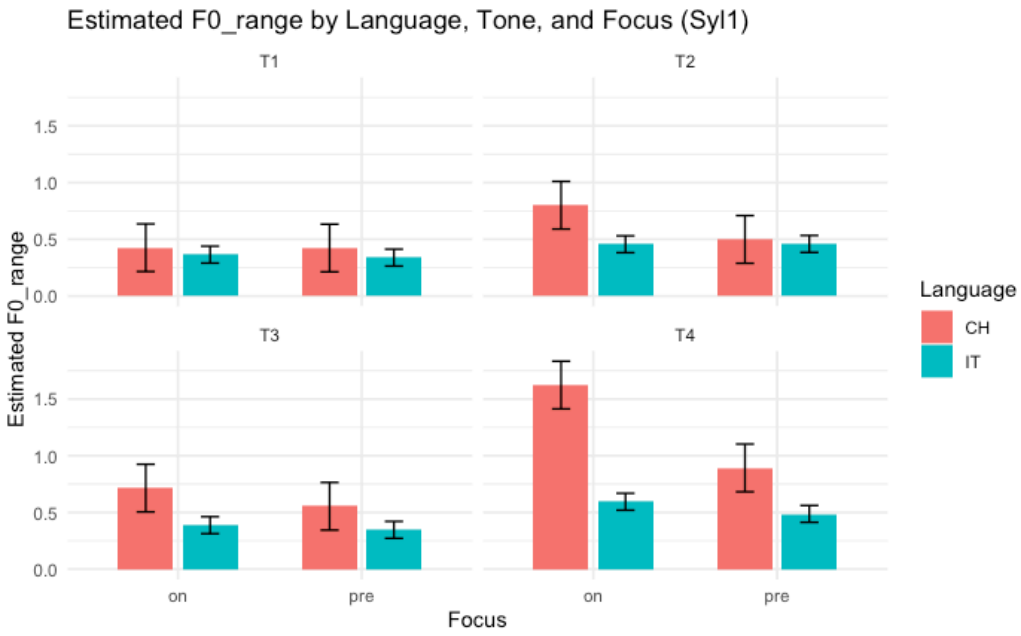


Figure 61 Estimated F0_range by Language, Tone, and Focus (Syl1)

As observable, CH robustly expand their pitch range under focus, especially in T4, whereas IT display much flatter range contours, with only minimal increases in pitch span under focus, and in some cases (notably T4), no significant expansion at all.

Taken together, these results confirm that F0_range functions as a sensitive and reliable correlate of prosodic focus, particularly in native speech. However, its effectiveness as a cue is compromised in L2 productions. While L1 speakers exhibit systematic pitch range expansion to enhance prominence, L2 speakers display a compressed dynamic range, particularly under conditions requiring lexical tone and prosody integration.

The significant three-way interaction underscores that focus realization is not merely a function of information structure, but rather emerges through a complex interplay between lexical and prosodic demands. For L2 learners, this interplay appears to pose a challenge, especially in T4, where range expansion is essential but under-realized. This suggests that L2 speakers have not yet acquired native-like prosodic flexibility, and that F0_range modulation may be an area of persistent difficulty in tonal L2 acquisition.

5.3.6 Italian learner subset

5.3.6.1 Comparing learner-factor models (Proficiency, Musicality, Grade)

To evaluate the extent to which individual learner characteristics modulate pitch production in disyllabic Mandarin phrases, a series of GAMMs were constructed and compared. The

modeling began with a baseline model incorporating only the interaction between Tone and Focus (hereafter referred to as the *TF* model). Subsequently, three additional models were developed by introducing one learner-specific variable at a time, namely Proficiency, Musicality, and Grade, to assess their respective contributions to explaining variation in F0 trajectories.

Each augmented model included a composite predictor variable created by crossing the individual difference factor with Tone and Focus (e.g., *PTF*, *MTF*, *GTF*), allowing for the examination of how pitch contours conditioned by Tone and Focus vary as a function of learner profile. All models also incorporated random smooths for Speaker and OtherTone, thereby controlling for individual speaker variability and tonal coarticulatory influences.

Model comparisons were performed using the *compareML()* function from the *itsadug* package. The results are summarized below:

Table 32 Comparison of learner-factor models for the focus analysis on Syl1

Comparison	AIC Difference	Best Model
TF vs. PTF	+27.23	PTF
TF vs. MTF	+57.00	MTF
TF vs. GTF	+283.87	GTF

All models incorporating individual difference variables yielded lower AIC values relative to the baseline model, indicating improved model fit. These results suggest that Proficiency, Musicality, and particularly Grade meaningfully enhance the model's capacity to account for variation in pitch realization across Tone.Focus contexts. Among the three, the *GTF* model yielded the largest AIC improvement, implying that educational stage may be the most influential factor in predicting how learners encode focus in these experimental conditions.

5.3.6.2 Interaction of Grade, Tone and Focus: developmental effects

Given that model comparisons identified the Grade.Tone.Focus (GTF) interaction model as providing the best explanatory fit, this model was subsequently refitted using the fREML estimation method, which is more robust and appropriate for inferential interpretation. Following this, smoothed F0 trajectories were plotted for each individual GTF level in order to examine how pitch contours unfolded over time across combinations of educational grade, tonal category, and focus condition.

In the broader CH-IT analysis on T1, both native and non-native speakers displayed higher F0 in on-focus conditions; however, Italian learners' T1 contours were notably more falling, diverging from the flatter, sustained pitch of native speakers. Interestingly, no significant group differences emerged in either focus condition, suggesting that IT broadly acquired the register-based aspects of T1, but not its fine-grained contour control (see § 5.3.1).

In this subset, more nuanced variation emerges by grade. Learners in the BA3 group produced lower-pitched and more steeply falling contours, particularly in on-focus contexts, suggesting a possible focus-marking strategy derived from negative phonological transfer. BA2 and MA1 learners exhibited relatively flatter – though still falling – contours with less pronounced decline, more closely approximating native-like patterns.

This contrast suggests that lower-level learners may rely on pitch cues influenced by negative L1 transfer – such as falling F0 – to signal prosodic prominence, whereas more advanced learners begin to internalize subtler, target-like realizations of the canonical tone shapes. This indicates that even at higher levels of proficiency, fine-grained tonal adjustments associated with focus marking remain underacquired, pointing to persistent difficulties in coordinating lexical tone with prosodic function in L2 Mandarin.

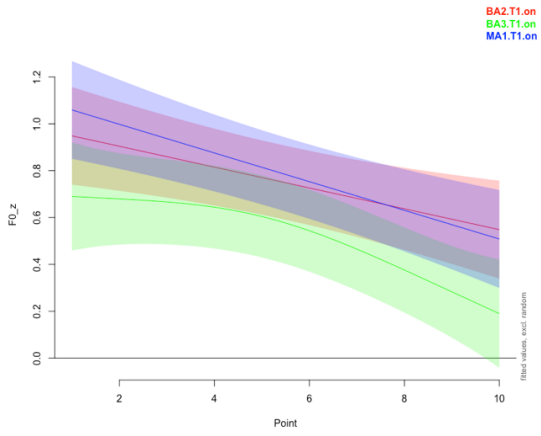


Figure 62 Tone 1 on-focus production by Grade (Syl1)

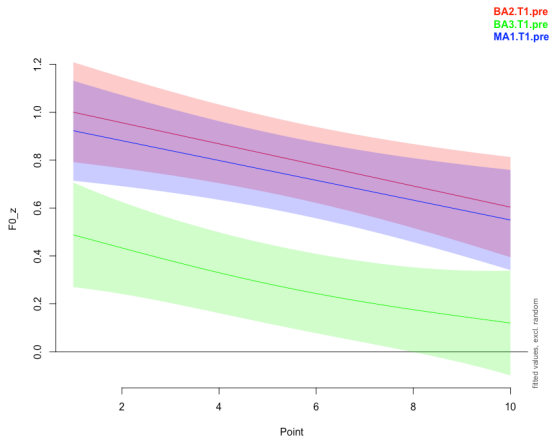


Figure 63 Tone 1 pre-focus production by Grade (Syl1)

The CH-IT comparison revealed that native speakers employed F0 lowering under focus for T2, likely to enhance contour shape (i.e., highlighting the rise), whereas Italian learners showed the opposite trend, raising F0 globally and flattening the rise. This suggests limited tone contour-prosody integration in the L2 group.

In the grade-level IT subset, MA1 learners distinguished themselves by producing falling-rising contours under focus, a pattern more closely aligned with the native-like realization of T2 in our experimental condition (see Fig. 64). By contrast, BA-level learners maintained elevated F0 values across the syllable, and in some cases – particularly BA2 pre-focus productions – exhibited falling contours that diverge from the canonical rising trajectory of T2 (see Fig. 65). While this deviation reflects incomplete acquisition of the tone's pitch shape, it may also hint at an incipient neutralization effect, suggesting that some learners are beginning to implement prosodic adjustments associated with information structure, albeit in a non-target-like manner. These patterns suggest a developmental shift: MA1 learners begin to implement tonal contour enhancement, whereas BA learners maintain simplified F0-raising or flattening strategies, especially in discourse contexts.

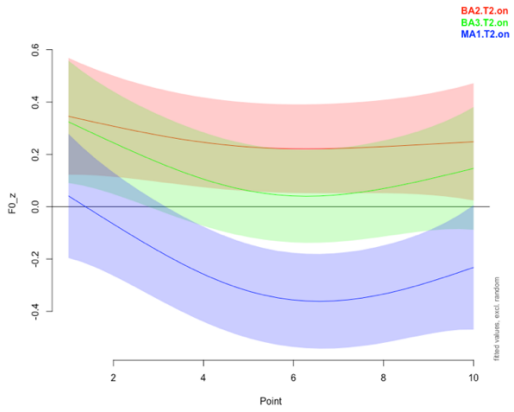


Figure 64 Tone 2 on-focus production by Grade (Syll)

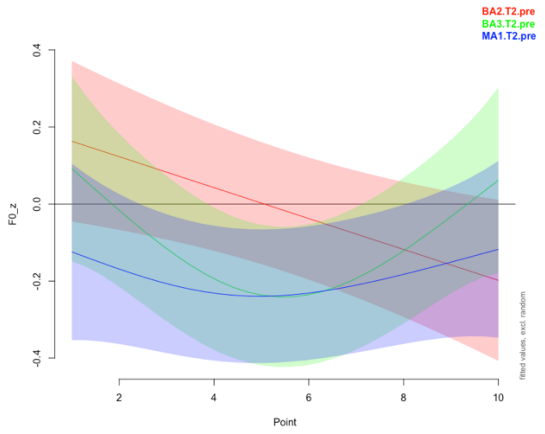


Figure 65 Tone 2 pre-focus production by Grade (Syll)

In the main CH-IT analysis, Italian learners consistently produced higher F0_z for T3 in both focus conditions, indicating difficulty targeting the low register. Focus exerted no significant effect on pitch in either group. Overall, L2 learners did not reproduce the canonical falling-rising contour associated with native realizations of T3.

The IT subset confirms this trend but adds new details. BA2 and BA3 learners failed to produce either falling or falling-rising shapes, particularly in pre-focus positions, confirming limited pitch modulation and tonal shape control. However, MA1 learners displayed slightly higher F0 offsets in pre-focus, possibly indicating anticipatory focus marking or tonal contour planning.

This suggests that L2 difficulty with T3 is persistent across educational levels, but some prosodic planning emerges in more advanced learners. The subset supports the broader findings while highlighting that pitch range compression and L1 intonational interference may diminish with experience.

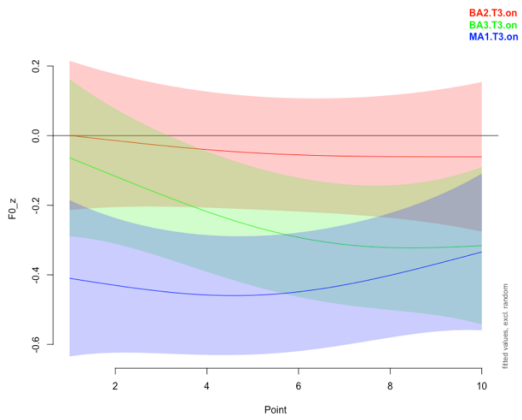


Figure 66 Tone 3 on-focus production by Grade (Syl1)

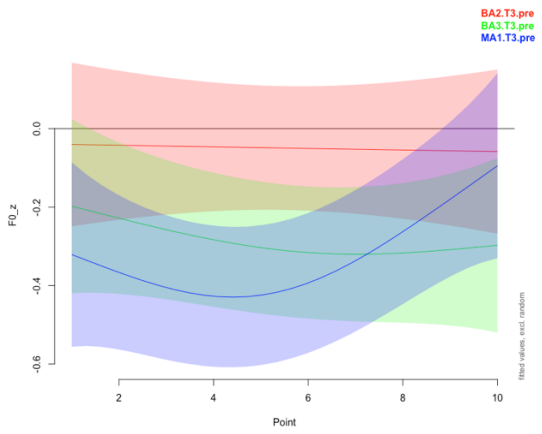


Figure 67 Tone 3 pre-focus production by Grade (Syl1)

In the CH-IT dataset, native speakers showed robust falling contours for T4 that were modulated by focus (steeper on-focus, more compressed pre-focus), while Italian learners exhibited flattened and less differentiated contours, especially under pre-focus, pointing to incomplete focus-tone integration.

In the IT subset, MA1 learners produced clearly falling contours under focus, with flattening in pre-focus, a pattern that closely aligns with native speakers, including evidence of tone neutralization pre-focus. In contrast, BA learners exhibited little variation across focus conditions, and BA3 learners even produced slight rising offsets in on-focus T4, diverging significantly from native patterns.

These results, depicted in Figg. 68 and 69, underscore that advanced learners not only improve tone shape accuracy but also acquire prosodic flexibility. BA learners, by contrast, exhibit intonational remnants of L1 and limited tonal modulation. The subset deepens the main CH-IT findings by demonstrating that T4 modulation, especially under focus, is a key diagnostic for tonal-prosodic integration in L2.

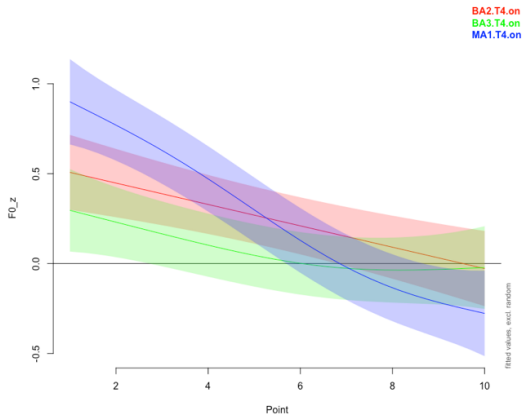


Figure 68 Tone 4 on-focus production by Grade (Syl1)

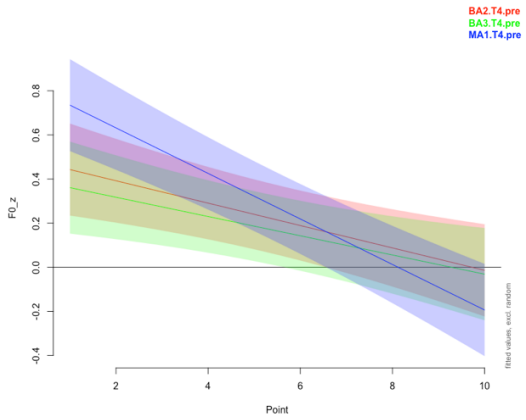


Figure 69 Tone 4 pre-focus production by Grade (Syl1)

5.3.7 Interim summary on Syl1

The Syl1 analysis offers a coherent account of how prosodic focus is realized across tones and speaker groups, with particularly revealing contrasts between CH and IT. Tone-wise, L1 and L2 productions diverge not simply in degree but in the strategies employed to mark focus.

For T1, both groups increased F0 under focus, but native speakers sustained a high, near-level trajectory to syllable offset, whereas learners produced a gradual F0 decline, indicating less stable maintenance of prominence. Notably, the grade-level subset reveals that lower-level learners exhibited more pronounced falling trajectories, suggesting a potential reliance on L1-derived pitch cues to mark prominence. In contrast, MA1 learners produced flatter contours, approaching more target-like realizations.

T2 yielded the most diagnostic asymmetries. CH showed a lowered F0 onset under focus, anchoring the rising contour more deeply and enhancing its contrastive shape. IT, by contrast, simply elevated the overall pitch, resulting in a higher but less dynamic rise. The subset analysis further refines this picture: MA1 learners produced falling-rising contours under focus, more closely aligning with native-like T2 shapes, while BA learners maintained elevated F0 across the syllable. Interestingly, BA2 learners' pre-focus productions even displayed falling contours, diverging from canonical T2 realization but potentially reflecting an early implementation of tone neutralization in pre-focus condition.

T3 revealed persistent group-level differences in pitch scaling: IT consistently produced higher F0 values and demonstrated limited modulation by focus, failing to achieve the low register characteristic of T3. This trend was echoed in the grade-level subset, where MA1 learners displayed higher F0 offsets pre-focus, possibly as a preparatory strategy, while BA-

level learners failed to produce clear falling-rising contours, underscoring a broader difficulty with low tonal anchoring.

T4 provided further insight into learners' challenges with pitch dynamics. Native speakers maintained a steep falling trajectory, even in pre-focus contexts, whereas learners' contours were shallower, and crucially, the descent was significantly flatter in pre-focus syllables, suggesting reduced flexibility in coordinating tonal shape with information structure. In the subset, MA1 learners' on-focus realizations exhibited a canonical fall, which flattened slightly in pre-focus. In contrast, BA learners showed minimal contour differentiation across focus conditions, with BA3 learners even displaying a slight rising offset on-focus, indicative of residual L1 intonational influence or incomplete tone-focus integration.

Complementary curve parameter analyses reinforce these findings. Mean F0 highlighted the clearest group separation for T2 and T3, where IT group's higher pitch levels point to insufficient register control. Slope analysis isolated the precise phonetic locus of prosodic mismatch, especially in T4's curtailed fall among IT in pre-focus. While both groups increased F0_max under focus – revealing that pitch peaks serve as a broadly accessible cue – F0_min and F0_range offered more discriminative power: learners maintained elevated pitch floors and exhibited attenuated range expansion, particularly in T2 and T4. These patterns were especially marked among BA learners, who demonstrated less pitch dynamism and tonal flexibility across conditions.

Taken together, these results on Syl1 suggest that native speakers encode focus in a tone-specific manner, flexibly adjusting both register and contour shape. In contrast, L2 learners – especially at lower levels – tend to adopt a more global register-based strategy, failing to implement the subtle contour manipulations required by tonal and prosodic integration. The primary bottleneck for the IT group thus lies not only in pitch height but in the interactive encoding of lexical tone and discourse prominence, a challenge that appears most acute for tones requiring large F0 excursions (e.g., low anchoring in T3, steep falling in T4). Importantly, MA1 learners show signs of progress toward target-like contour shaping.

5.4 Second-syllable analysis

To examine the prosodic realization of focus on the second syllable of disyllabic target phrases, a GAMM was fitted using F0_z as the dependent variable. The model tested the three-way interaction between Language (L1 Mandarin vs. L2 Italian), Lexical Tone (T1-T4), and Focus condition (on-focus vs. post-focus). Smooth terms were included for each

Language.Tone.Focus combination, allowing the modeling of time-varying pitch trajectories over the course of the syllable; additionally, the model incorporated random smooths for Speaker and OtherTone, which captured inter-speaker variability and effects of tonal coarticulation, respectively[18].

The parametric portion of the model provides insight into global differences in mean F0 height across model conditions. Using native speakers T1 on-focus (CH.T1.on) as the reference level, several significant effects emerged.

In terms of Language effect, IT exhibited significantly lower F0_z than CH in the reference condition, with a drop of 0.31 Hz ($p < .001$). Focus effect implied that post-focus conditions triggered a marked global lowering of F0 across both groups (estimate = -0.90, $p<.001$), in line with expectations of PFC. As for the tone effect, T2 and T3 were associated with significantly lower average F0 than T1 (-1.49 and -1.14, respectively, $p < .001$), reflecting their canonical phonetic realizations. Furthermore, several interaction terms were highly significant. In particular, the Language.Tone.Focus term (LangIT.ToneT2.Focuspost = -0.76, $p < .001$) showed that Italian learners failed to maintain pitch height under T2 in post-focus contexts. Similarly, the Language.ToneT4 and Language.ToneT4.Focuspost terms were both significant ($p < .001$), indicating flattened contours and reduced post-focus pitch suppression in IT T4 productions. These findings demonstrate that F0 height patterns vary systematically by tone and focus, and that IT exhibit L2-specific deviations, particularly under prosodic conditions that require fine control of pitch scaling.

In addition to modeling mean F0 height, the GAMM included smooth terms over normalized time (Point) for each model condition. These smooths allowed the model to capture the internal structure of pitch contours within the syllable, revealing how tonal shapes are dynamically realized and how they vary by language and focus.

Most on-focus conditions exhibited significant smooth terms ($p<.001$), with estimated effective degrees of freedom (edf) ranging from ~1.0 to ~5.1. This variation reflects the complexity of the tone contour in each context.

For T2, both CH and IT speakers produced complex contours. CH.T2.on had an edf of 3.64, while IT.T2.on exhibited even greater complexity, likely reflecting L2 over-articulation or hypercorrection in rising tones. For T1, the contrast was striking. While CH.T1.on revealed a

[18] The model showed stable convergence under fREML optimization using discrete estimation, and diagnostics on the basis dimension (k = 10) revealed no evidence of undersmoothing. All k-index values were approximately 0.98, and all associated p-values were non-significant or marginal, confirming that the model captured the appropriate degree of complexity in pitch contour shape.

low-complexity but significant smooth (edf = 1.74, p<.001), IT.T1.on had a non-significant smooth term (edf = 1.00, p = .108), suggesting a flat or underspecified pitch contour. Post-focus conditions were often marked by reduced smoothness or non-significance, in line with PFC effects. For instance, both CH.T1.post and IT.T1.post had non-significant or weak smooth terms, indicating the suppression of tonal movement after the focus domain.

Table 33 Smooth terms across Syl2 conditions

Condition	edf	F	p-value	Interpretation
CH.T1.on	1.74	12.69	< .001	Slight rising movement
IT.T1.on	1.00	2.58	0.108	Flat / no tonal shaping
CH.T2.on	3.64	13.13	< .001	Canonical rising tone
IT.T2.on	4.91	31.55	< .001	Complex rising curve
CH.T3.post	2.44	13.34	< .001	Moderate curvature preserved
IT.T4.post	3.51	17.19	< .001	Flattened T4 post-focus

Overall, these smooth terms offer a detailed view of contour dynamics, showing that L1 speakers preserve tonal shape across prosodic contexts, while L2 speakers often flatten or distort the expected pitch trajectories, particularly in post-focus conditions and for T1 and T4.

To complement the GAMM results, we conducted EMMs comparisons within each tone using the emmeans package. These pairwise comparisons clarified how Focus interacts with Language within each Tone category. Below, the relevant contrasts for each target tone are reported.

5.4.1 Tone 1

For T1, the data revealed a robust focus effect among CH group. Parametric estimates indicated that on-focus syllables were produced with significantly higher normalized F0 values compared to post-focus ones (CH.on - CH.post = +0.94, p < .0001; see also § 5.4.5.1 on Mean F0 analysis results). This reflects a canonical T1 focus-marking strategy involving F0 elevation on the focused constituent, followed by PFC, a prosodic pattern widely attested in Mandarin (Xu, 1999; Chen, 2010). The GAMM-derived contours reported in Fig. 70 further corroborate this interpretation: the on-focus curve exhibits a mild rising trend, while the post-focus curve remains flat and compressed, underscoring the dynamic use of pitch to mark information structure.

In contrast, L2 learners of Mandarin displayed only a subdued version of this pattern. While a directionally similar trend was observed – slightly higher F0 in on-focus than post-focus positions – the extent of modulation was considerably attenuated. Most notably, the on-focus F0 in L2 productions was significantly lower than that of native speakers (CH.on - IT.on = +0.35, p = .0005). Moreover, the predicted F0 trajectory in L2 on-focus productions lacked the rising movement seen in native speech, suggesting that learners relied more on static pitch levels than dynamic pitch contours to encode prosodic prominence for T1 on syl2. Additionally, PFC was minimal and statistically non-significant in L2 speech, indicating limited implementation of PFC as a focus marking cue. These results suggest that while Italian learners may possess a coarse awareness of focus-related F0 modulation, their ability to deploy native-like prosodic strategies remains constrained, particularly for tones characterized by level pitch targets like T1.

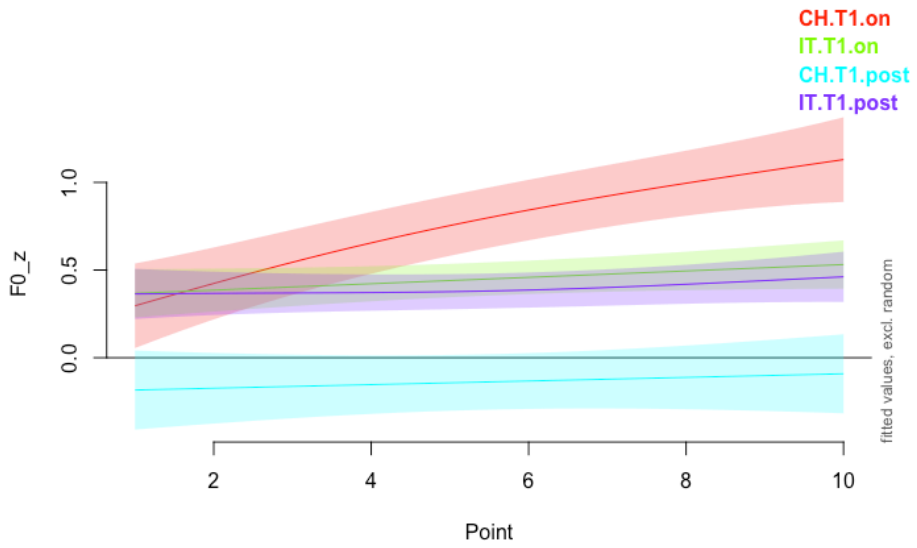


Figure 70 Tone 1 production by Language and Focus

5.4.2 Tone 2

For T2, an inverse pattern was observed when comparing CH and IT in terms of pitch height: L2 speakers displayed atypically elevated F0 values in both on-focus and post-focus positions, higher than native speakers on-focus productions.

Specifically, L2 learners produced significantly higher F0_z in on-focus syllables than native speakers (CH.on-IT.on = -0.30, p = .0236). This pattern may reflect a hyper-articulation

strategy, whereby learners overproduce pitch height in an attempt to enhance the perceptual salience of the rising contour associated with T2. This overemphasis is consistent with learner reliance on segmental or tonal cues in isolation, at the expense of broader prosodic integration, as demonstrated by the absence of a clear F0 distinction between focus conditions in the L2 group. In contrast to native speakers, whose productions were characterized by focus-induced pitch suppression in post-focus positions, L2 learners did not exhibit significant PFC ($p = .12$), indicating a limited implementation of post-focal prosodic restructuring.

Visual inspection of the GAMM-predicted curves (see Fig. 71) further substantiates this pattern. The language group difference is most pronounced in the initial portion of the syllable, where learners start from a higher pitch baseline. Native speakers, however, showed clear focus-related modulation toward the final portion, where the rising trajectory of T2 is undershoot under post-focus conditions – an effect absent in learner productions (see § 2.3.2). These findings suggest that learners may lack the fine-grained temporal control and prosodic flexibility to encode focus in coordination with T2 tonal target on syl2.

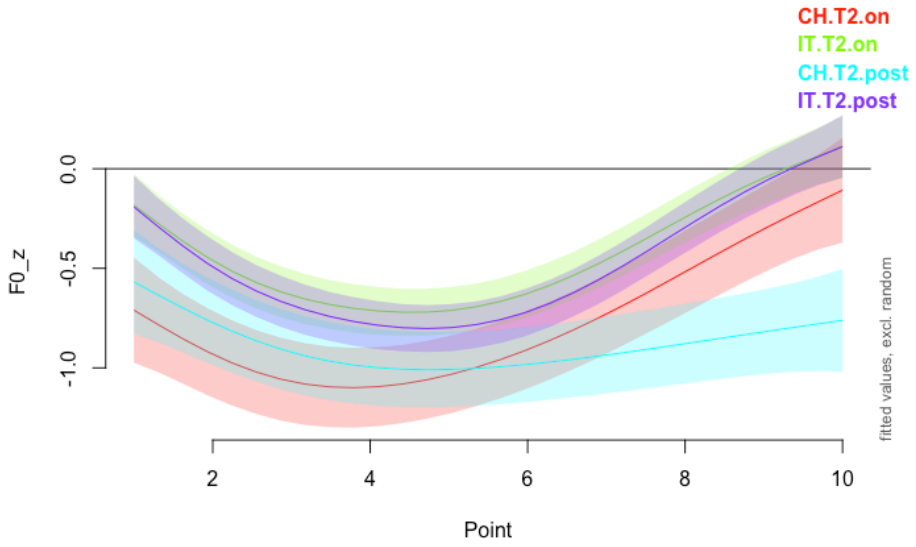


Figure 71 Tone 2 production by Language and Focus

5.4.3 Tone 3

In T3, no statistically significant differences were observed between L1 and L2 groups across focus conditions. The contrast between CH.on and IT.on was negligible (+0.03, $p = .99$),

suggesting comparable pitch behavior across both language groups, apart from a more explicit rising after the dipping portion in IT productions. This finding implies that T3, which involves a low-dipping contour, may be less susceptible to focus-driven pitch modulation or that its acoustic properties make it inherently less marked for prosodic emphasis.

Despite the absence of statistically significant group-level differences, the GAMM-derived F0 trajectories reveal subtle yet informative qualitative distinctions. L2 speakers tended to produce T3 with largely overlapping contours across focus conditions, maintaining a consistent citation-like falling-rising shape. This pattern suggests an adherence to the canonical full-T3 contour, reflecting learners' orientation toward phonological targets at the lexical level. In contrast, native speakers' productions were more variable and did not include the final rise typically associated with the full T3. Instead, both focus conditions exhibited a "half-T3" form, a well-documented variant in natural speech, likely resulting from tonal reduction and contextual assimilation (Xu, 1997; see also § 2.2.4).

Crucially, however, a prosodic distinction was still evident in native productions: in on-focus contexts, the post-dip portion of the contour reached a relatively higher F0 than in post-focus contexts – indicating a subtle pitch enhancement aligned with focus marking. This prosodic modulation was absent in the L2 group, whose F0 trajectories remained consistent across conditions (see also § 5.4.5.1 on Mean F0 analysis results). These findings suggest that, while learners may succeed in approximating the phonological contour of T3 in isolation, they often lack the fine-grained temporal control required for context-sensitive prosodic adjustments, particularly in the late-syllable domain where information structure effects typically emerge. Taken together, these findings imply that T3 may serve as a point of early convergence between native and non-native speakers, precisely because its focus-marking potential is reduced by its phonological form. The lack of post-dip distinction in L2 productions, however, indicates incomplete integration of prosodic cues and supports the broader claim that learners often prioritize segmental and tonal targets at the expense of discourse-level pitch modulation.

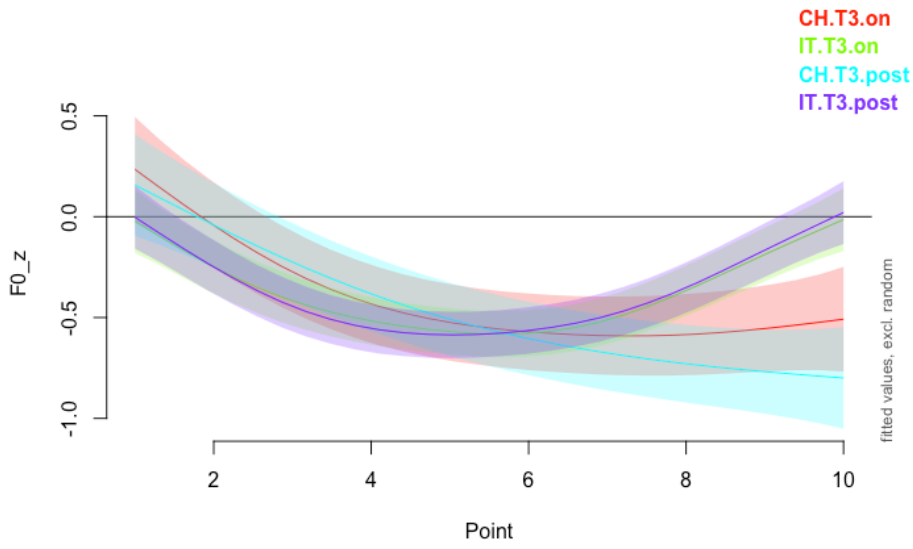


Figure 72 Tone 3 production by Language and Focus

5.4.4 Tone 4

Similarly to Syl1 results, the most striking cross-linguistic differences on Syl2 emerged for T4. CH exhibited both high F0_z in on-focus contexts and a significant decrease in post-focus syllables (CH.on-CH.post = +0.63, p<.0001), reflecting strong and systematic focus marking through pitch. In contrast, IT produced significantly flatter contours, with no meaningful difference between on- and post-focus conditions (IT.on-IT.post = +0.02, p=.95). The difference in on-focus F0 between CH and IT was both large and highly significant (CH.on-IT.on = +1.12, p<.0001), indicating a failure to achieve the tonal target or to appropriately mark prosodic prominence (see Fig.73).

The absence of focus-related modulation in the realization of IT T4 suggests a phonological underspecification of falling tones and underscores the particular challenges faced by Italian learners in producing tonal contours that require precise pitch descent in phrase-final position. Previous research in this work (see § 4) has already demonstrated the absence of an adequately convex-like falling contour in isolated target productions; what emerges as novel in the present data is that this difficulty persists across the intonational phrase and remains unaffected by focus. In other words, T4 is produced in a manner closely resembling its citation form, without systematic variation as a function of focus. This pattern stands in contrast to native speaker productions, which not only exhibit a more phonologically distinct convex-falling T4 contour but also modulate its realization according to focus marking.

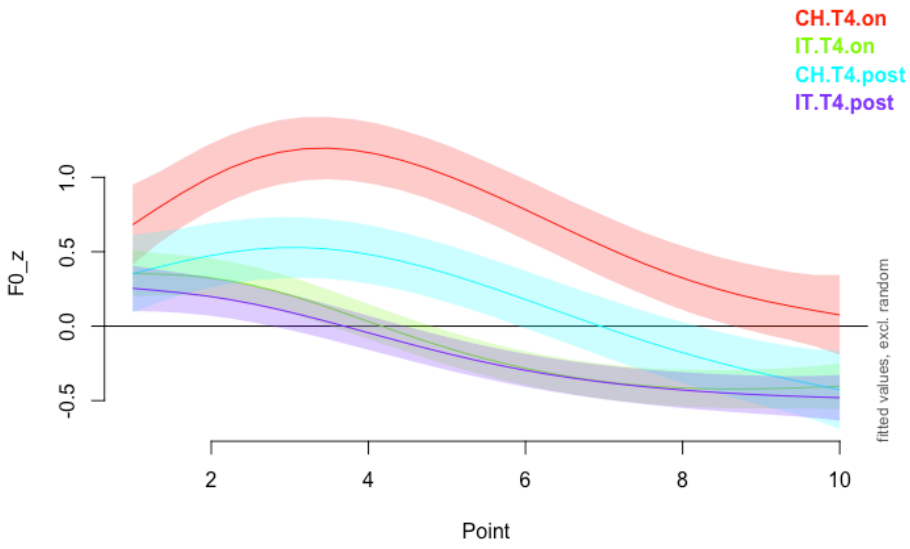


Figure 73 Tone 4 production by Language and Focus

5.4.5 Analysis on Curve Parameters for Syl2

5.4.5.1 Mean F0

To investigate whether Mean F0 in the second syllable systematically reflects prosodic focus, and whether this effect is modulated by tone and L1 background, a GLMM was fitted. The fixed structure included a full three-way interaction between Language, Lexical Tone, and Focus condition. Speaker and OtherTone were included as random intercepts to account for repeated measures and tonal coarticulation effects, respectively. Model comparison using log-likelihood ratio tests confirmed the necessity of both random intercepts (both $p < .0001$).

The fixed-effects structure of the model revealed a robust three-way interaction between Language, Tone, and Focus ($F(3,2940.93)=3.17$, $p=0.023$). This suggests that the way in which pitch height responds to focus is not uniform across tones, and critically, that this tonal sensitivity differs between native and non-native speakers.

All three main effects – Language, Tone, and Focus – were significant or marginally significant, as were all two-way interactions. Notably, the Language.Focus interaction was highly significant.

Results points to a general failure by IT to implement post-focus lowering, a well-established prosodic cue in native Mandarin speech. Tab. 34 below summarizes the key fixed-effect estimates from the final model:

Table 34 Mean F0 model key fixed-effect estimates

Effect	Estimate	p-value	Interpretation
Intercept (CH, T1, on)	+0.303	0.028	Baseline pitch height for CH speakers in T1.on
LangIT	+0.138	0.318 (n.s.)	L2 slightly higher pitch (not significant)
ToneT2	-1.208	*** < .001	T2 significantly lower than T1
ToneT3	-0.493	** 0.004	T3 lower than T1
ToneT4	+0.008	0.962 (n.s.)	No difference from T1
Focuspost	-0.842	*** < .001	Strong post-focus lowering
LangIT.ToneT2	+0.476	** 0.008	IT raise T2 relative to CH
LangIT.ToneT4	-0.311	.084 (.)	Some flattening of T4 in IT
LangIT.Focuspost	+0.824	*** < .001	IT suppress post-focus lowering
ToneT2.Focuspost	+0.690	** 0.004	Focus raises T2 pitch (counterintuitive interaction)
ToneT3.Focuspost	+0.715	** 0.003	Similar effect as T2 for T3
LangIT.ToneT2.Focuspost	-0.628	* 0.014	IT flattens pitch in T2 post-focus
LangIT.ToneT3.Focuspost	-0.713	** 0.005	Similar effect as T2 in T3

These results highlight the central role of tone in shaping prosodic strategies, and the clear divergence between CH and IT in how pitch is used to mark focus. While native speakers produced a consistent pattern of pitch lowering in post-focus positions across tones, IT frequently fail to modulate pitch appropriately – particularly in T2 and T3, where focus-driven pitch suppression is either absent or reversed.

The post-hoc EMMs plotted below (see Fig. 74) illustrate the observed interaction effects. CH display the expected PFC as Mean F0 is lower in post-focus syllables. In contrast, IT show minimal or null focus-induced pitch adjustment.

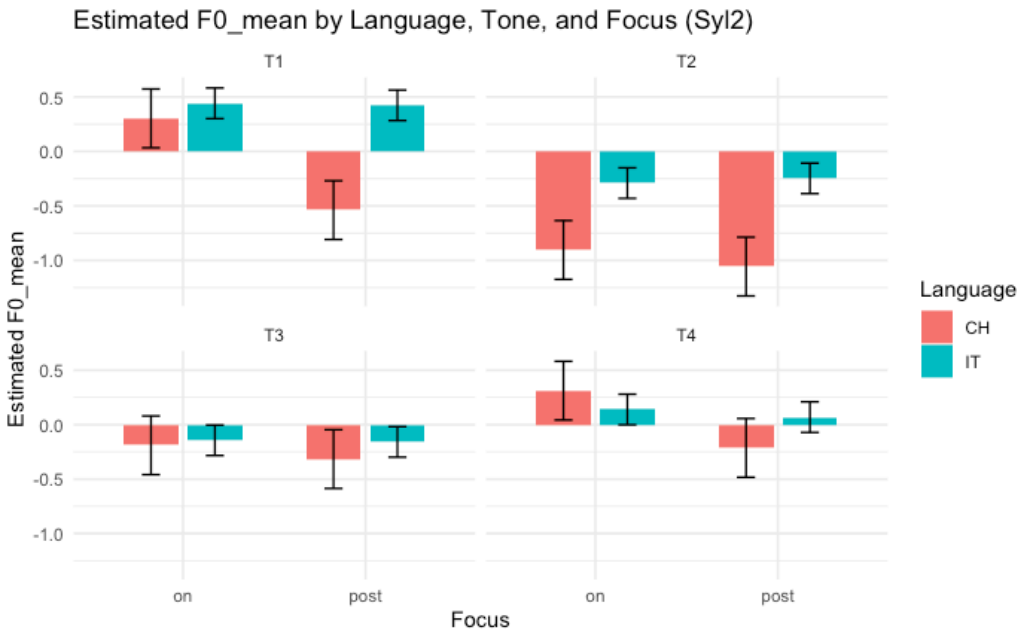


Figure 74 Estimated Mean F0 by Language, Tone, and Focus (Syl2)

These findings offer important insights into L2 prosody in a tonal language. First, Mean F0 is an adequate correlate of focus for native speakers, especially in phrase-final syllables, where PFC serves as a reliable cue to information structure.

Second, although Italian learners approximate tone height distinctions, they fail to modulate pitch height systematically across focus conditions. The observed three-way interaction – particularly in T2 and T3, where IT diverge sharply from CH pitch behaviors – suggests that tone-specific prosodic control remains underdeveloped. Indeed, IT do not exhibit the prosodic flexibility required to align lexical tone with discourse-level focus marking.

This has important implications for second language phonological development, suggesting that pitch height manipulation under prosodic constraints represents a persistent challenge for learners from non-tonal L1 backgrounds.

5.4.5.2 F0 slope

To determine whether F0_slope on the second syllable (Syl2) serves as a cue for prosodic focus and whether it varies systematically across languages and tonal categories, a GLMM was fitted. The dependent variable was F0_slope, calculated as the coefficient of a linear regression over the time-normalized F0 points for each syllable.

Random effects were specified for Speaker and OtherTone. Likelihood ratio tests confirmed the necessity of both grouping factors[19].

Initial model comparisons yielded evidence that Focus, along with all interactions involving Focus (e.g., Tone.Focus, Language.Focus, Language.Tone.Focus), did not improve the model fit and could be eliminated without significant loss. This outcome suggests that, unlike in the first syllable, F0_slope does not serve as a robust correlate of focus in the second syllable, for either native or non-native speakers. Consequently, the final model retained only Language, Tone, and their two-way interaction.

Importantly, neither Language nor Tone had significant main effects after accounting for their interaction. This implies that pitch slope differences are not uniform across speakers or tones, but instead emerge in specific Language-Tone combinations.

The EMMs demonstrate that while both Mandarin and Italian speakers adhere to general tonal contours – rising in T2, dipping in T3, and falling in T4 – they differ in how sharply these slopes are realized.

For T2, IT exhibited significantly steeper rising slopes compared to native speakers. This may reflect hyper-articulation or prosodic overcompensation, possibly due to L2 speakers' effort to clearly mark a rising tone.

For T3, a similar hyperarticulation was found, with IT producing sharper pitch excursions than CH, especially in the final portion of the predicted curve, targeting the full falling-rising contour taught for the T3 citation form.

For T4, in contrast, IT produced a less steep falling slope compared to native speakers, indicating prosodic smoothing or target undershoot. This may be due to limited control of falling contours, which are more complex to execute with appropriate pitch range compression.

These patterns are visualized in Fig. 75, which plots the estimated slope values across the four tones by language group. The figure illustrates both over-shooting in T2 and T3 and under-shooting in T4 by IT relative to CH norms.

[19] Removing either term resulted in a significant decrease in model fit (Speaker: $\chi^2 = 27.54$, $p < .001$; OtherTone: $\chi^2 = 7.11$, $p = .008$), validating their inclusion.

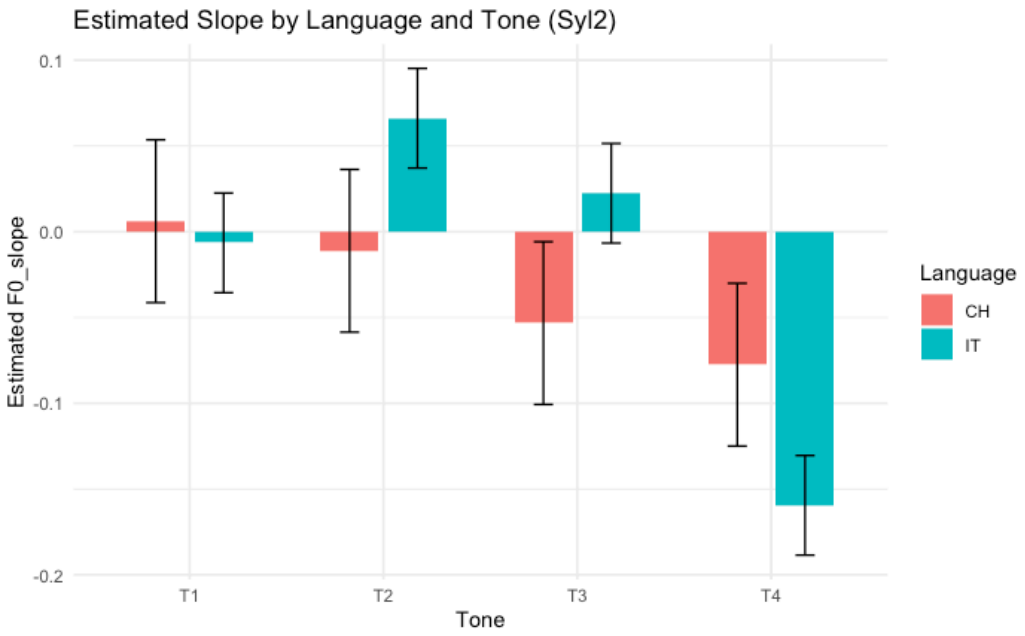


Figure 75 Estimated Slope by Language and Tone (Syl2)

These findings suggest that F0_slope in the second syllable does not contribute meaningfully to focus marking in either native or non-native speech. This contrasts with the behavior observed in Syllable 1, where slope played a clearer role in prosodic prominence. The lack of a Focus effect may reflect post-nuclear deaccentuation or pitch reset, which often flattens pitch movement in phrase-final positions.

Nevertheless, the Language.Tone interaction reveals systematic divergences in how slope is realized across tones. IT do not modulate pitch slope uniformly across tone categories, and instead demonstrate a tone-specific strategy, characterized by either hyper-articulation (on T2, T3) or prosodic flattening (on T4). This asymmetry suggests that pitch slope in L2 Mandarin may be governed more by lexical-level tone targets than by discourse-level structure, such as focus.

5.4.5.3 F0 max

To investigate how maximum pitch (F0_max) functions in the expression of prosodic focus and lexical tone in the second syllable, a GLMM was fitted. The model included Language, Tone, and Focus as fixed factors, along with their full factorial interaction structure. Speaker and OtherTone were incorporated as random intercepts to control for speaker-specific variation and tonal coarticulation effects, respectively.

Model comparison via likelihood ratio tests confirmed the importance of including both random intercepts[20].

Following model simplification through backward elimination, the final model retained all main effects and two two-way interactions: Language.Tone and Language.Focus. Importantly, the three-way interaction (Lang.Tone.Focus) and the Tone.Focus interaction did not significantly improve model fit and were therefore excluded. This suggests that the effect of focus on peak pitch does not differ systematically across tone categories in a language group-specific manner.

The main effect of Focus revealed a significant lowering of F0_max in post-focus conditions ($\beta = -0.582$, $p < .001$), consistent with PFC trend. This effect was most robust among L1 speakers, who displayed consistent suppression of peak pitch in post-focus environments.

Crucially, however, the Language.Focus interaction indicates that Italian learners exhibited a reduced magnitude of post-focus pitch suppression, pointing to a less consistent implementation of PFC. While both groups modulate pitch height to mark focus, L2 speakers did so less dynamically, resulting in a narrower contrast between on- and post-focus maxima.

A further Language.Tone interaction underscores that peak pitch modulation varies not only by focus condition but also by tone category in a language group-dependent trend. In particular, the interaction terms for T3 and T4 revealed significantly lower F0_max values in IT compared to CH productions. This suggests pitch target undershoot for tones that require complex contouring, especially falling or dipping tones, which demand wider pitch excursions and are typically more difficult for non-native speakers to acquire and control.

The EMMs of F0_max across Language and Focus conditions are visualized in Fig. 76, illustrating that while CH display a clear drop in maximum pitch post-focus across tones, IT produce a more compressed pitch maximum with attenuated focus effects, especially in T3 and T4.

[20] Removing Speaker led to a significant deterioration in model fit ($\chi^2(1) = 60.92$, $p<.001$), and OtherTone also significantly contributed ($\chi^2(1) = 24.15$, $p < .001$), indicating that peak pitch is shaped not only by individual speaker differences but also by the tonal environment in which a syllable occurs.

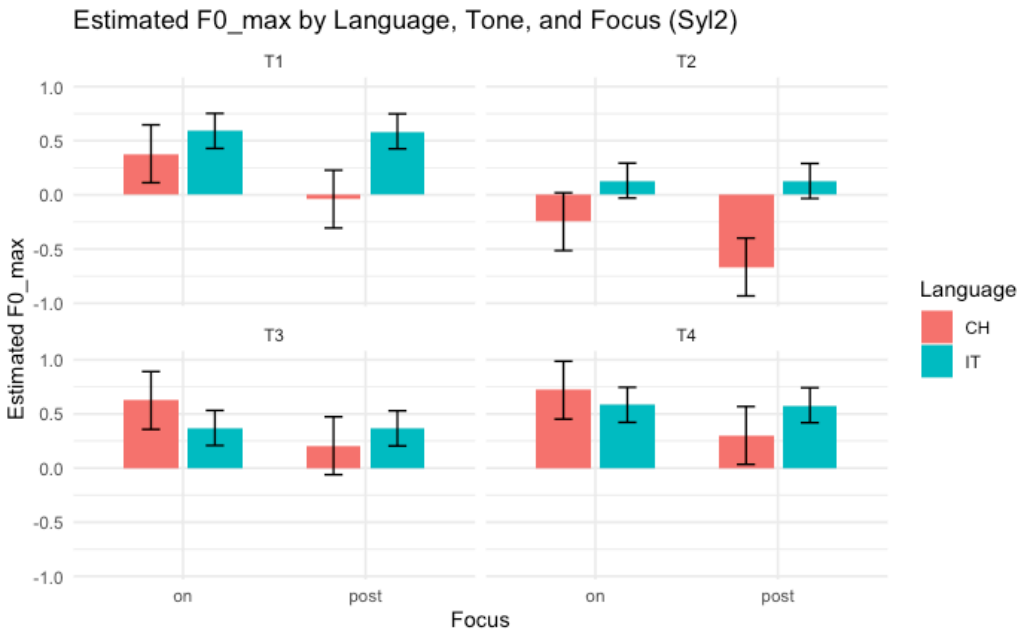


Figure 76 Estimated F0_max by Language, Tone, and Focus (Syl2)

The results confirm that F0_max serves as a reliable prosodic correlate of focus in L1 Mandarin, particularly in syllable-final positions where PFC is expected. For Italian learners, focus-driven modulation of F0_max is highly reduced, indicating a less flexible prosodic system that may reflect limited acquisition of PFC in L2 Mandarin.

Furthermore, the tone-dependent divergence – with L2 speakers significantly underproducing peak values for T3 and T4 – proves that contour tones with falling trajectories pose particular challenges for L2 learners of Mandarin. This suggests that L2 speakers' difficulty in modulating pitch range is compounded in contexts where both lexical tone and information structure require coordinated pitch movement.

Taken together, these findings highlight a key aspect of L2 prosody: while learners may approximate native-like behavior in general pitch scaling (e.g., raising pitch under focus), their ability to compress or suppress pitch post-focus, and to do so differentially across tonal categories, remains limited. Thus, F0_max is a partially robust focus marker in L2 speech, but its effectiveness is compromised by tonal undershoot and a reduced prosodic contrast system, especially in tones with greater pitch movement demands.

5.4.5.4 F0 min

To assess how minimum pitch (F0_min) varies as a function of Language background, Tone, and Focus conditions in the second syllable of disyllabic Mandarin phrases, a GLMM was

constructed with full factorial interactions. The model included Language, Tone, and Focus as fixed effects, along with random intercepts for Speaker and OtherTone[21].

Critically, the analysis revealed a significant three-way interaction between Language, Tone, and Focus, demonstrating that the effect of focus on F0_min varies across tones, and that this variation is modulated by language background. The fixed effects summary (Tab. 35) highlights the key patterns:

Table 35 F0_min model fixed effects summary

Effect	Estimate	*p*-value	Interpretation
LangIT	+0.306	0.067 (.)	IT have marginally higher pitch floor overall
ToneT2	-1.176	< .001 ***	T2 has substantially lower F0_min than T1
Focuspost	-0.827	< .001 ***	Strong post-focus pitch floor lowering
LangIT.Focuspost	+0.822	< .001 ***	IT exhibit reduced pitch floor suppression post-focus
ToneT3.Focuspost	+1.004	< .001 ***	T3 shows increased pitch in post-focus
LangIT.ToneT3.Focuspost	-0.978	0.0017 **	IT fails to suppress pitch floor in T3 post-focus
LangIT.ToneT2.Focuspost	-0.696	0.025 *	IT also undercompresses T2 in post-focus

These interactions demonstrate that CH consistently lower the pitch floor in post-focus contexts, particularly for T2 and T3, both of which involve low tonal targets as part of their phonological specification. In contrast, IT exhibit a markedly reduced ability to compress the pitch floor, especially in these same tones, suggesting incomplete acquisition of tone-prosody integration.

The EMMs of F0_min by Tone, Focus, and Language are visualized in Fig.77 below. The pattern is clear: for T2 and T3, CH display notably lower F0_min in post-focus conditions, whereas IT F0_min values remain elevated and largely invariant across focus conditions. The interaction effects are much less pronounced in T1 and T4, in line with the non-significant interaction terms for those tones.

[21] Model comparison via likelihood ratio tests confirmed the necessity of both random effects: excluding Speaker significantly worsened model fit ($\chi^2(1) = 47.63$, $p < .001$), as did removing OtherTone ($\chi^2(1) = 12.84$, $p < .001$).

This contrast is particularly revealing: while native speakers dynamically lower the pitch floor to enhance PFC – thereby contributing to focus marking through pitch range expansion – Italian learners fail to enact this cue reliably, especially when the lexical tone itself demands low pitch realization.

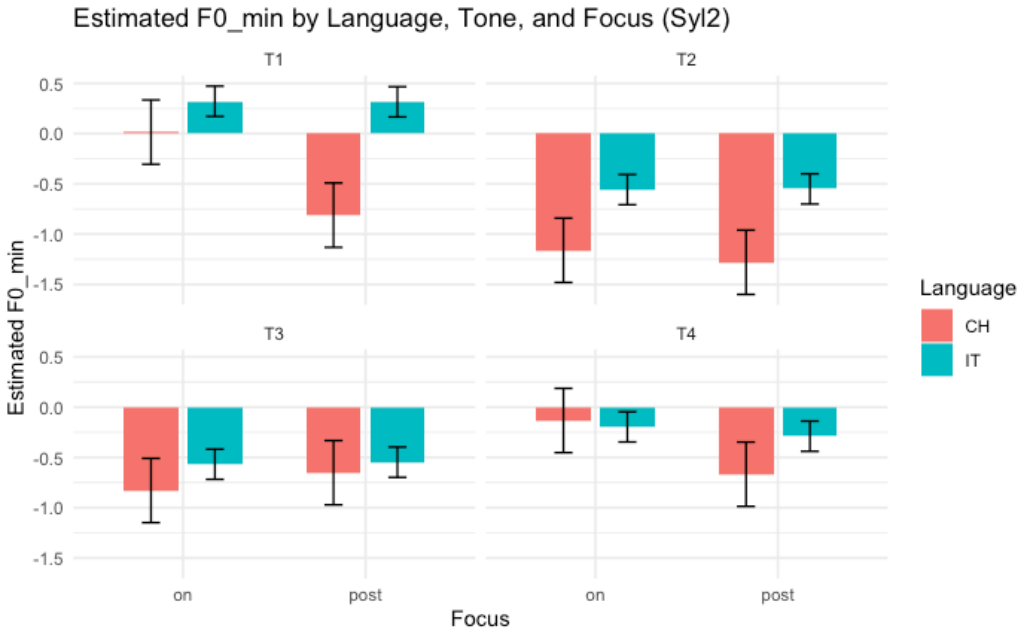


Figure 77 Estimated F0_min by Language, Tone, and Focus (Syl2)

The findings strongly suggest that F0_min is a sensitive and meaningful correlate of focus marking in L1 Mandarin, particularly when combined with tonal categories that require low pitch realization. However, Italian learners appear to underutilize pitch floor lowering, resulting in a more narrowed pitch range and weakened prosodic contrast between focus and post-focus contexts.

The significant Lang.Focus and Lang.Tone.Focus interactions indicate that this limitation is not only general but also tone-specific. In tones with inherently low targets – such as T2 and T3 – IT fail to implement sufficient PFC, possibly due to limited control over low-pitch register or interference from L1 prosodic patterns. Importantly, the positive LangIT.Focuspost effect coupled with negative LangIT.Tone.Focuspost terms confirms that IT not only compress pitch less overall, but fail to adapt this compression to the tonal identity, a key element of native-like prosodic modulation.

The analysis of F0_min in syllable-final positions reveals a robust interaction between tone, focus, and language background, demonstrating that learners exhibit persistent

undercompression of the pitch floor, especially in tones requiring low-pitch targets. These findings reinforce the view that minimum pitch is an essential acoustic marker for both lexical and information structure cues, and that pitch floor manipulation should be a focus of pedagogical intervention in the prosodic training of L2 Mandarin learners.

5.4.5.5 F0 range

In order to assess the extent to which speakers modulate pitch span within the second syllable of disyllabic Mandarin utterances, we analyzed F0_range, defined as the difference between the maximum and minimum normalized F0 values per syllable. This measure provides an index of prosodic dynamism, capturing the degree of pitch excursion speakers produce during the realization of lexical tones, and is particularly informative for evaluating prosodic flexibility in focus conditions and across speaker groups.

A GLMM was fitted with Language, Tone, and Focus as fixed effects, and with Speaker and OtherTone as random intercepts. Model comparison via likelihood ratio testing revealed that Speaker accounted for significant variance in F0_range ($\chi^2(1) = 32.05$, $p < .001$), while OtherTone contributed no explanatory power ($\chi^2(1) \approx 0$, $p = 1$), and was therefore excluded. This confirms that individual variability among speakers meaningfully shapes pitch span, whereas immediate tonal context exerts minimal influence in this particular acoustic domain.

Subsequent fixed-effects testing and backward model reduction led to a final model that retained only main effects of Language and Tone. Neither Focus nor any two-way or three-way interactions significantly improved model fit.

The final model yielded the following fixed-effect statistics:

Table 36 F0_range fixed-effect statistics

Effect	F	*p*-value	Interpretation
Language	4.01	0.049 *	Small but significant difference between CH and IT groups
Tone	54.24	< .001 ***	Robust variation in pitch span across tones
Focus	0.006	n.s.	No significant focus-related modulation of F0_range
Lang.Tone	0.76	n.s.	No tone-specific differences in IT pitch span
Lang.Focus	0.75	n.s.	Focus structure did not modulate CH-IT group differences
Tone.Focus	1.25	n.s.	Focus-related pitch span modulation not tone-dependent

These results underscore two key findings. First, IT exhibit significantly compressed pitch range relative to CH, regardless of tonal identity or prosodic focus. Second, Focus structure

does not significantly modulate F0_range in either group, suggesting that pitch span is not actively recruited as a correlate of focus in the phrase-final position.

A bar plot based on EMMs illustrates the F0_range patterns across tones and languages (Fig. 78). The data reveal a consistent flattening of pitch range in IT across all tonal categories. Both groups display the expected tonal hierarchy: T3, with its dipping contour, exhibits the widest pitch span, followed by T4 and T2. T1, as a level high tone, naturally shows the narrowest pitch range. Crucially, the absence of significant interactions indicates that IT do not modulate pitch range in a tone-specific manner, contrasting with the more differentiated patterns observed in CH.

The lack of any significant effect of Focus (or interaction with Language or Tone) suggests that, at least within the final syllable of disyllabic phrases, F0_range is not systematically employed to mark prosodic focus. This contrasts with earlier findings in Syllable 1, where F0_range indicate clearer sensitivity to focus structure.

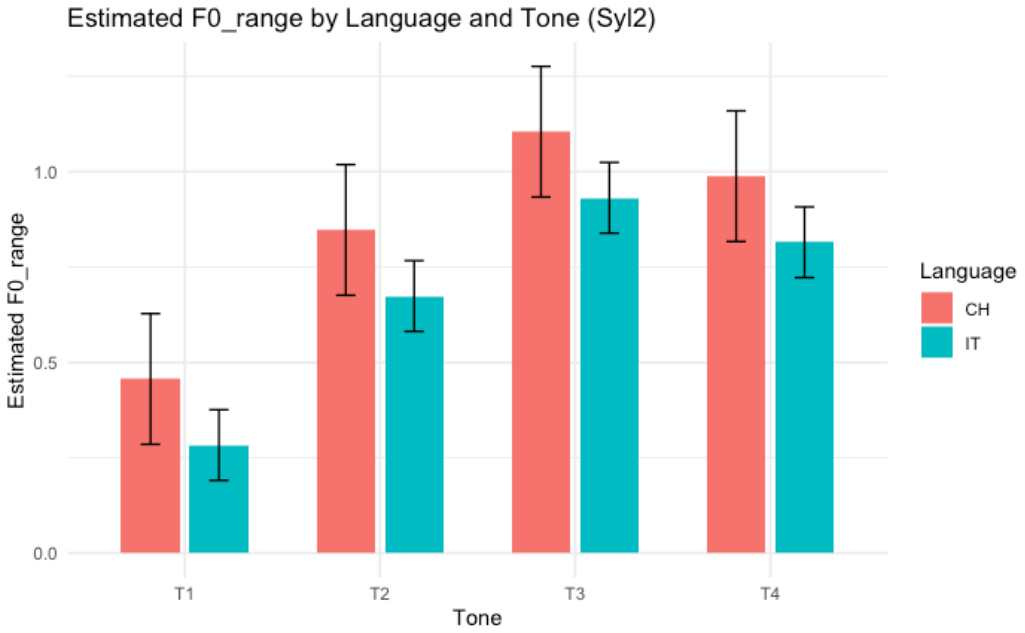


Figure 78 Estimated F0_range by Language and Tone (Syl2)

The finding that IT consistently undershoot native-like pitch span, regardless of tonal category or information structure, provides compelling evidence for a general flattening of prosodic contours in non-native Mandarin speech.

While tonal identity still exerts a significant influence on pitch span – reflecting phonological requirements of the Mandarin tonal system – learners appear to lack the ability

to deploy this parameter contrastively, especially in ways that would enhance the communicative marking of focus. The result is a compressed prosodic profile, which may contribute to both reduced intelligibility and lowered naturalness in L2 tonal productions.

The analysis of F0_range in Syllable 2 reveals a clear main effect of language background, with learners demonstrating globally reduced pitch span across all tones. Unlike native speakers, they fail to leverage pitch range to mark information structure, nor do they adapt their span modulation in a tone-specific way. Although T3 retains the widest range in both groups, reflecting its phonological contour, the prosodic contrast between focus conditions is largely neutralized in L2 speech.

These findings reinforce the interpretation that F0_range is a valuable acoustic index of prosodic proficiency, particularly in tone languages. The consistent compression of pitch range in L2 Mandarin productions suggests that greater attention to pitch span control – especially in expressive or contrastive contexts – may be beneficial in both theoretical modeling of L2 tone learning and in pedagogical intervention.

5.4.6 Italian learner subset

5.4.6.1 Comparing learner-factor models (Proficiency, Musicality, Grade)

To evaluate the extent to which individual learner characteristics modulate pitch realization in the second syllable (Syl2) of disyllabic phrases, we implemented a series of GAMMs. These models were fitted to the Italian learner subset of the dataset, focusing on the interaction between Tone and Focus as the core condition structure.

We first constructed a baseline model that included the interaction between Tone and Focus (TF model) as a fixed effect, along with random factor smooths for Speaker and OtherTone (i.e., preceding syllable tone), accounting for inter-speaker variability and local tonal context. To explore the role of individual difference variables, we then created three additional models by extending the grouping variable to include Proficiency (PTF), Musicality (MTF), and Grade (GTF), crossed with Tone and Focus. This allowed us to test whether learner-level factors improved model fit by explaining additional variance in the pitch trajectories. All models were fitted using maximum likelihood (method = "ML"), and compared using compareML() from the itsadug package.

Table 37 Comparison of learner-factor models for the focus analysis on Syl2

Comparison	AIC Difference	Best Model
TF vs. PTF	+0.23	PTF
TF vs. MTF	+0.78	MTF
TF vs. GTF	+266.80	GTF

While both Proficiency and Musicality yielded small improvements in model fit, the Grade.Tone.Focus model (GTF) showed a dramatically better fit, with an AIC difference of +266.80 over the baseline. This indicates that educational stage is a more robust predictor of how learners encode tonal and prosodic information in Mandarin Syl2, compared to proficiency level or musical aptitude as measured in this study. These findings are consistent with those obtained in the Syl1 analysis, where Grade also emerged as the strongest predictor of tonal realization under focus.

5.4.6.2 Interaction of Grade, Tone and Focus: developmental effects

As Grade emerged as the strongest predictor of tonal realization under focus, the GTF model has been refitted with fREML. This analysis complements the broader cross-linguistic GAMM results presented for Syl2 by examining developmental differences among L2 speakers, with particular attention to the effects of academic level (BA2, BA3, MA1) and focus condition (on-focus vs. post-focus).

In the production of T1 on the second syllable, BA3 learners exhibited a slight rising trajectory under focus, a pattern that approximates the marginal rising offset characteristic of native speaker realizations in similar conditions. In contrast, BA2 and MA1 learners produced smoother, more compressed contours in on-focus contexts, with MA1 speakers additionally displaying elevated F0 across the syllable. In post-focus position, BA2 learners demonstrated a mild rising tendency, likely attributable to phrase-final boundary effects rather than intentional post-focal modulation.

When considered in relation to the larger dataset including native speakers, these findings highlight persistent L2-specific deviations from native-like prosodic encoding. Native Mandarin speakers were found to maintain a high, near-level F0 trajectory under focus, followed by subtle pitch suppression in post-focus contexts – a pattern consistent with canonical PFC. Crucially, none of the learner groups systematically implemented this dynamic contrast. Instead, their post-focus productions remained relatively flat and marginally lower, but without statistically reliable evidence of PFC.

These patterns suggest that learners across Grade levels experience difficulties in integrating information-structural cues such as focus into tonal realization also for tone categories like T1, which lack salient pitch movement. The lack of dynamic F0 shaping indicates that learners tend to use global pitch height as a cue to prominence, instead of employing the subtle temporal modulations observed in native prosody.

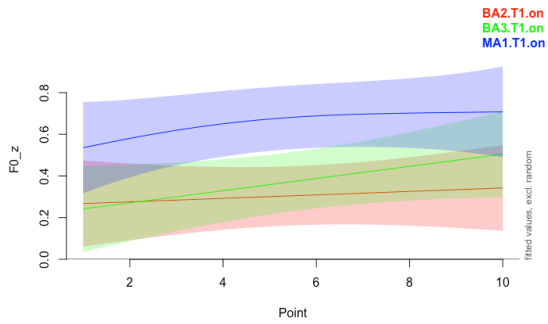


Figure 79 Tone 1 on-focus production by Grade (Syl2)

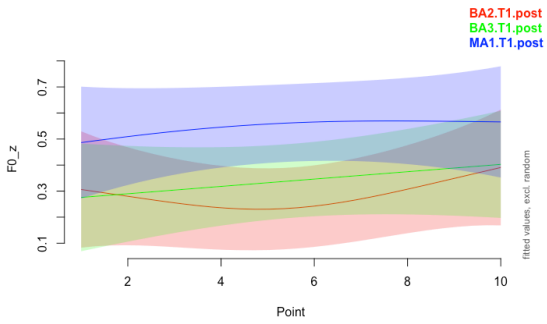


Figure 80 Tone 1 post-focus production by Grade (Syl2)

T2 exhibited relatively uniform behavior across learner subgroups, with minimal differentiation attributable to either focus condition or grade level. BA3 learners, in particular, consistently produced elevated F0 values in both on-focus and post-focus contexts, with the pitch increase becoming most prominent from the mid-syllable onward. Given the absence of focus-related contrast, this elevation is unlikely to reflect intentional focus marking; rather, it likely represents a general over-raising of the pitch target independent of information structure.

These findings align closely with the results from the broader analysis including native speakers, which showed that learners tend to over-articulate T2 in on-focus positions – producing higher-than-native F0 peaks – and failed to implement PFC effectively. The lack of tonal distinction between focus conditions in learner productions suggests limited prosodic flexibility and a reliance on hyperarticulated pitch cues to approximate the canonical rising contour of T2.

The uniform absence of PFC across groups further underscores a broader difficulty in integrating tonal and pragmatic functions in L2 Mandarin, particularly for tones that rely on pitch scaling over time rather than discrete contour shapes.

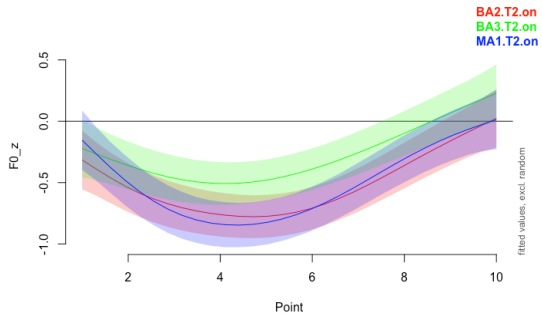


Figure 81 Tone 2 on-focus production by Grade (Syl2)

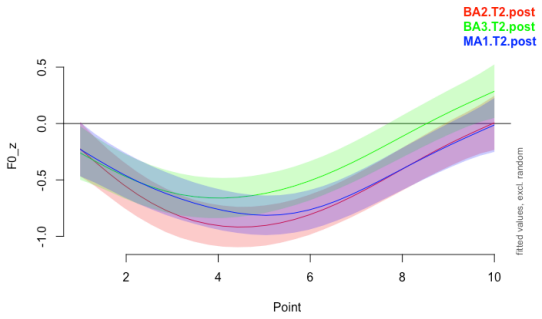


Figure 82 Tone 2 post-focus production by Grade (Syl2)

Learner productions of T3 were marked by a consistent adherence to the full dipping contour characteristic of the citation form, with this tendency particularly evident in the BA2 and MA1 groups. This suggests that these learners primarily relied on phonological representations acquired in isolation, rather than adapting their tonal productions in response to discourse-level cues. BA3 learners, while likewise exhibiting minimal focus-driven modulation, produced overall higher F0 values across both on-focus and post-focus conditions – consistent with the elevation trend observed in their T2 productions.

These results corroborate findings from the broader GAMM including native productions, which indicated that native Mandarin speakers systematically reduce the post-dip rise of T3 in post-focus contexts, producing a compressed "half-T3" form – a prosodic reduction strategy also associated with PFC. In contrast, L2 learners retained the full contour across focus conditions, providing little to no evidence of PFC.

The subset model thus underscores a broader pattern: while Italian learners are able to approximate the phonological shape of T3, they exhibit limited capacity to modulate this shape according to discourse-pragmatic constraints. This points to a persistent difficulty in integrating tonal form with information structure, and highlights T3 as a locus where phonological knowledge is not yet fully mapped onto prosodic function in L2 speech.

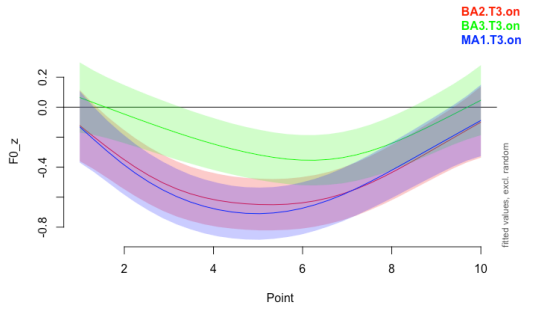


Figure 83 Tone 3 on-focus production by Grade (Syl2)

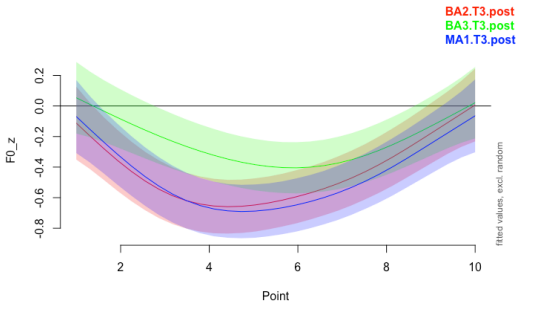


Figure 84 Tone 3 post-focus production by Grade (Syl2)

Among the four lexical tones, T4 demonstrated the most pronounced developmental trajectory across learner groups. MA1 learners exhibited notably steeper falling contours in both on-focus and post-focus conditions, closely resembling the canonical citation form. This suggests a more advanced degree of tonal realization, likely reflecting increased phonetic control at the segmental level. In contrast, both BA learner groups exhibited markedly flatter or only shallowly falling contours across focus conditions – a pattern consistent with their Tone 4 realizations in isolated citation contexts (see § 4) –, suggesting a persistent reliance on underspecified tonal targets even when additional prosodic demands are present. BA2 learners exhibited an overall higher F0 offset in the post-focus position, further diverging from the expected PFC pattern.

These findings are consistent with those observed in the larger model, which revealed that learners systematically under-realized the falling trajectory of T4 and failed to implement the pitch lowering typically associated with PFC. The subset analysis corroborates this result: while MA1 learners appear to approximate the tonal target more faithfully in its phonological form, none of the learner groups demonstrated robust prosodic modulation across focus conditions.

Taken together, the results suggest that only the most advanced learners (MA1) exhibit signs of improved tonal production for T4, yet even they show limited evidence of integrating prosodic structure with lexical tone. This pattern highlights the complexity of acquiring falling tones in L2 Mandarin, particularly in discourse contexts where fine-grained pitch control is required. It also reinforces the broader claim that successful acquisition of tone contour alone does not guarantee native-like prosodic behavior, especially when that tone must interact with information-structural demands such as focus.

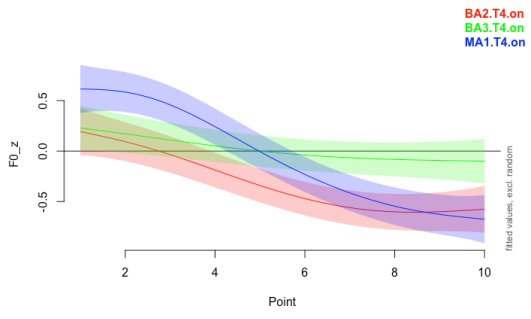


Figure 85 Tone 4 on-focus production by Grade (Syl2)

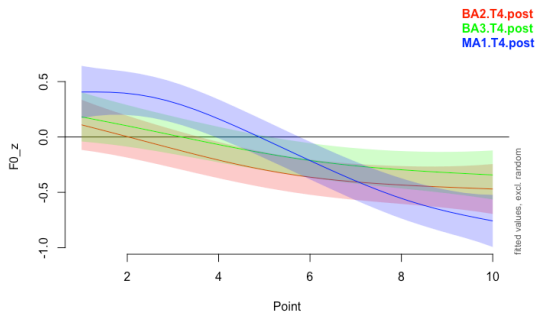


Figure 86 Tone 4 post-focus production by Grade (Syl2)

5.4.7 Interim summary on Syl2

The analysis of the second syllable (Syl2) in disyllabic phrases reveals a clear tone-dependent asymmetry between CH and IT in how prosodic focus is encoded at phrase boundaries. While the broader GAMM, including the native control group, confirmed robust PFC in native speech – manifesting as both global F0 lowering and contour neutralization – learner productions consistently exhibited attenuated or absent versions of these cues. The analysis of the Italian subset, using Grade.Tone.Focus (GTF) interactions, extends these observations by identifying developmental trajectories in L2 tonal-prosodic integration, highlighting variation across BA2, BA3, and MA1 learners.

In the broader model, T1 served as a relatively unmarked tone with respect to contour shape, yet native speakers still exhibited dynamic modulation: a high, level trajectory under focus and a compressed, flat contour in post-focus contexts. Italian learners, by contrast, demonstrated a simplified version of this strategy – raising F0 under focus to a lesser extent and failing to reliably lower pitch post-focus. The GTF subset analysis refines this picture. BA3 learners displayed a marginally rising trajectory in on-focus position, partially converging with the native pattern, though without sufficient PFC. BA2 and MA1 learners produced smoother, less differentiated contours, with MA1 showing a higher register overall but no more dynamic modulation. These results suggest that even advanced learners continue to rely on static register shifts rather than fine-grained temporal modulations to mark focus when the tone lacks contoural salience. The absence of PFC across grades reinforces the broader conclusion that level tones like T1 pose integration challenges for learners, precisely because they demand subtle adjustments rather than overt contouring.

T2 emerged as the most diagnostic tone in the broader model: native speakers produced a deeper rise under focus and flattened the contour post-focus (cfr. tone-target undershoot), while L2 learners significantly over-raised pitch in both conditions and failed to implement PFC. This was interpreted as hyper-articulation – an attempt to preserve the perceptual salience of the rising contour in the absence of prosodic nuance. Learner subset results confirmed this interpretation. BA3 learners consistently produced elevated F0, particularly from the mid-syllable onward, in both focus conditions. Since this elevation was uniform, it is unlikely to signal focus marking. BA2 and MA1 exhibited similar over-raising with minimal variation across conditions, reinforcing the interpretation that learners default to phonological over-specification at the expense of discourse-sensitive modulation. The lack of within-group differentiation suggests that even increased educational level (MA1) does not ensure target-

like PFC implementation for T2, underscoring persistent difficulty at the tone-prosody interface.

The main analysis accounting for T3 showed that native speakers produced a reduced "half-T3" form post-focus, suppressing the post-dip rise in line with PFC conventions. In contrast, learners tended to maintain the full contour across conditions, reflecting a reliance on citation-form representations rather than context-sensitive adjustments.

The GTF model subset findings echoed this pattern. BA2 and MA1 learners reproduced the full dipping contour regardless of focus, while BA3 learners also demonstrated minimal modulation, albeit with slightly elevated F0. The persistence of the full T3 shape across grades suggests that learners prioritize phonological fidelity over prosodic adaptability. Crucially, none of the learner groups implemented the prosodic compression observed in native speech. This indicates that T3, despite being well approximated in phonetic form, remains resistant to prosodic conditioning in L2 acquisition – a revealing dissociation between form knowledge and functional deployment.

T4 provided the clearest site of divergence between learners and native speakers. The broader model showed that native speakers produced steep falling contours on-focus, followed by a marked lowering in post-focus syllables. Learners, by contrast, flattened the fall and failed to implement PFC, yielding near-neutralization of the on/post focus contrast.

The learner subset analysis identified a clear developmental trend. MA1 learners produced relatively steep falling contours in both conditions, more closely approximating the citation target, though still lacking dynamic focus marking. In contrast, both BA groups produced markedly flatter contours across conditions. BA2 learners also produced an elevated offset in post-focus position, deviating further from native-like behavior. These findings confirm that while advanced learners (MA1) may approximate segmental tone targets, integration of focus-related prosody remains limited. The shallowness of BA productions and lack of focus modulation suggest that lower-level learners may rely on static, underspecified tonal realizations, especially discourse contexts.

5.5 Discussion

The combined analyses of the first (Syl1) and second (Syl2) syllables provide a comprehensive picture of how Italian L2 learners of Mandarin encode prosodic focus in disyllabic phrases. Across both positions, the findings converge on the same core result: learners generally approximate the phonological targets of Mandarin lexical tones in isolation,

but systematically underperform in integrating those targets into a discourse-conditioned prosodic system. This deficiency is especially evident in post-focus contexts, where native Mandarin speakers deploy robust PFC through both pitch lowering and contour simplification. Learners, by contrast, tend to rely on more global register adjustments, resulting in attenuated or absent on/post focus contrasts.

Tone-specific comparisons across Syl1 and Syl2 reveal that native speakers mark focus not simply by raising F0 but by reshaping tonal contours in a manner appropriate to each tone.

For T1, both groups exhibit pitch raising under focus; however, native speakers sustain a high, level F0 contour until offset and produced systematic compression in pre- and post-focus regions. In contrast, learners, especially those at earlier stages of instruction, exhibit gradually falling or flattened F0 contours and provide little evidence of pre- or post-focus pitch suppression.

T2 emerged as the most diagnostic tone across both syllables. In both positions, native speakers anchor the rising contour more deeply under focus, creating a clear information-structural contrast. Learners, however, over-raise pitch in both conditions – particularly BA3 in Syl2 – without differentiating focus contexts. This hyper-articulation, observed consistently across grades, signals a reliance on phonological over-specification rather than prosodic flexibility. Even MA1 learners, who produced a falling-rising contour in Syl1 on-focus tokens, failed to implement native-like PFC in Syl2. This indicates that the tone-prosody interface for T2 remains underdeveloped even at advanced stages.

For T3, native speakers across both syllables reduce the post-dip rise in post-focus contexts, producing a compressed “half-T3” consistent with natural speech reduction. Italian learners instead retain the full fall-rise contour across focus conditions, reflecting their adherence to citation forms learned in isolation. The subset analysis confirmed this pattern: BA2 and MA1 learners produced the canonical dipping shape regardless of focus, while BA3 learners showed slightly elevated F0 but no greater prosodic modulation. This consistency across syllables underscores a broader developmental bottleneck: learners master T3’s phonological shape but fail to remap it onto contextually appropriate prosodic realizations.

T4 consistently produced the largest divergence between groups across both syllables. Native speakers exhibit steep on-focus falls and strong post-focus lowering, maintaining tone identity while dynamically marking focus. Learners flatten the fall and largely neutralize the on/post contrast. The subset analysis revealed a clear developmental trend: MA1 learners produced relatively steeper falls approximating the citation target, while both BA groups

displayed markedly flatter contours. BA2 learners even produced elevated offsets in post-focus position, a pattern inconsistent with canonical PFC. The replication of this pattern across Syl1 and Syl2 highlights T4 as a locus of persistent difficulty, particularly in controlling low pitch and steep declinations under discourse constraints.

Curve-parameter analyses for both syllables converge on the same mechanisms. Mean F0 consistently indicated a three-way interaction between Language, Tone, and Focus, with native speakers lowering post-focus height across tones and learners suppressing or reversing this effect, especially in T2 and T3. F0_max increased under focus for both groups, but the Lang.Focus interaction indicated that learners' peak suppression post-focus was smaller and their peaks were undershot for T3 and T4. F0_min emerged as the most sensitive index of native PFC: L1 speakers depressed the pitch floor most in T2 and T3 post-focus, while L2 speakers maintained elevated minima, producing compressed spans – a limitation especially marked in phrase-final position (Syl2). Finally, F0_range was globally smaller in the IT group and showed little or no focus-related modulation in Syl2, indicating reduced prosodic dynamism precisely where native Mandarin exhibits its strongest PFC.

The Grade.Tone.Focus subset models for both syllables clarify that Grade is a stronger predictor of tonal-prosodic integration than Proficiency or Musicality. While MA1 learners exhibit signs of improved phonetic control (e.g., steeper T4 falls, emerging falling-rising T2 contours), they still provide limited evidence of dynamic focus marking, particularly in post-focus contexts. BA-level learners default to oversimplified or over-articulated contours, relying on static, underspecified tonal realizations across discourse conditions. This developmental pattern, replicated across Syl1 and Syl2, reinforces the view that phonological contour knowledge precedes prosodic plasticity: learners first acquire tone shapes but only later begin to remap them onto context-sensitive prosodic functions.

Taken together, these results refine our understanding of the tone-intonation interface in L2 Mandarin. Italian learners' primary bottleneck lies not in identifying or reproducing tonal categories, but in implementing low targets, steep declinations, and focus-conditioned span control in intonational phrases – this is especially true at phrase boundaries, where PFC is most robust in native speech. Their reliance on L1 prosodic strategies, such as global F0 raising and narrower pitch spans, leads to prosodic simplification that blunts tonal contrasts and weakens the on/post focus distinction.

Across both Syl1 and Syl2, the evidence converges: Italian L2 learners have largely acquired the citation forms of Mandarin tones but not their discourse-conditioned prosodic modulation.

Native speakers mark focus by coordinating peaks and floors in a tone-specific manner; learners instead rely on register adjustments that produce higher but less informative contours and compressed spans. This tone-prosody integration gap is most acute for low-target and falling tones, particularly under post-focus conditions. Addressing these specific control problems promises the largest gains in intelligibility and pragmatic adequacy for L2 Mandarin.

Pedagogically, these findings argue for moving beyond static tone templates and incorporating dynamic, discourse-level prosody into instruction. Training should explicitly target: 1) post-focus compression as floor-lowering, not just peak control; 2) maintaining steep, context-sensitive falls for T4 at phrase boundaries; 3) tempering T2 hyper-raising to allow native-like neutralization (i.e., tone target undershoot) in post-focus contexts.

By coupling tone-specific contour practice with focus scenarios and feedback on pitch span management, instruction shall foster the tone-intonation integration that characterizes native production.

6 Prosodic encoding of Sentence type in L2 Mandarin: how sentence type, focus and tone interact

The present analysis examined tone production on the second syllable of disyllabic target phrases as a function of focus (on-focus vs. post-focus conditions) and sentence type, specifically contrasting statements with echo questions. Speech data were elicited from a cohort of 42 adult Italian learners of Mandarin Chinese, alongside a control group of 10 native Mandarin speakers. For the methodological reasons outlined in Chapter 3, T3 was excluded from the present analysis. Accordingly, the final dataset comprised 24,960 observations, including T1,T2, and T4. Fundamental frequency (F0) was converted to z-score normalized values (F0_z) per speaker to allow for direct comparison across individuals, thereby controlling for between-speaker variability in pitch range and register.

6.1 Research questions and hypotheses

This study examines how Italian university learners of Mandarin Chinese (MC) encode tone, focus, and sentence type interactively in phrase-final position. Specifically, it addresses the following research questions (RQs):

RQ1. To what extent do learners exploit tonal contour – rather than pitch register – to signal sentence type (e.g., echo questions vs. statements) in utterance-final position, potentially disregarding lexical tonal specifications in favor of boundary-tone-like prosodic cues, in line with tendencies observed in their L1 prosodic system?

RQ2. In what ways do learners' prosodic encoding strategies converge with or deviate from native Mandarin intonation patterns, and to what extent are these differences predicted by Grade, Proficiency, and Musicality?

It is hypothesized that Italian learners of Mandarin rely more heavily on pitch contour modulation, particularly in phrase-final contexts, than on absolute pitch register to mark interrogativity. For example, learners may realize a canonical falling tone (T4) with a rising or level contour to encode a question, thereby overriding its lexical tonal specification. This tendency is expected to be most pronounced among lower-proficiency learners, although contour-based strategies are anticipated to persist to some extent even among advanced speakers. This may reflect a broader instructional gap, as university-level Mandarin programs in Italy typically lack explicit training in prosodic or intonational features of the language.

Consequently, learners may draw on strategies rooted in L1 prosodic systems (e.g., Italian), leading to systematic cross-linguistic influence in L2 tonal prosody.

6.2 Dataset Overview

The dataset consists of annotated syllables extracted from the primary experimental task (see § 3.3) and includes productions from both L1 Mandarin speakers and L2 Mandarin learners. For the present analysis, productions of both declarative and interrogative utterances were included, with the scope restricted to second-syllable tokens. Each syllable was uniquely identified and annotated for a range of linguistic and prosodic variables, as outlined below:

- SylID: Unique identifier for each syllable;
- Speaker: Individual speaker identifier;
- Lang: Language group (CH = native Mandarin speakers; IT = L2 Mandarin learners);
- Tone: Lexical tone of the target syllable (T1, T2, T4);
- OtherTone: Lexical tone of the preceding syllable;
- Focus: Focus condition (on-focus, post-focus);
- F0: Raw fundamental frequency (Hz), sampled at 10 time-normalized points per syllable;
- F0_z: Speaker-wise z-score normalized F0 values.

In addition, a set of derived parameters was computed to capture the level, shape, and dynamism of the F0 contour, including mean F0, F0_max, F0_min, F0_range, and F0_slope.

6.3 Modelling the Interaction of Language, Sentence Type, and Focus

To explore how sentence type (statement vs. question) and focus condition (on-focus vs. post-focus) interact with language background (CH vs. IT), a GAMM was fit using *bam()* function with a three-way interaction modeled via smooth terms across time (Point). A new interaction factor *LSF* (Lang.S.Type.Focus) was created to facilitate condition-specific smooths[22].

[22] The model converged successfully under fREML estimation (fREML = 34089), with a scale estimate of 0.880 and adjusted $R^2 = 0.139$. All basis dimensions were adequate (k-index ≈ 1.02, all $p > 0.89$), indicating no undersmoothing. Although the model explained 14.7% of the deviance, this level of fit is reasonable given the intricate interaction structure, the breadth of condition levels, and the exclusion of the Tone variable, which is likely to contribute additional explanatory power in subsequent analyses.

Parametric contrasts relative to the reference level (CH.question.on) revealed highly significant differences between most LSF levels ($p < 0.001$). For instance, Italian learners in question-on condition (IT.question.on) showed significantly lower F0_z compared to native speakers (estimate = -0.554, $p < 0.0001$). In statement-on and question-post conditions, Italian learners also produced significantly lower F0_z. The CH.statement.post condition had the lowest predicted F0_z, indicating strong PFC among native speakers.

Smooth terms point to strong time-varying effects for Italian learners in all conditions except statement.on, suggesting dynamic pitch movement during intonational encoding. In contrast, native speakers exhibited flattened or non-significant smooths in several conditions – likely reflecting more categorical pitch implementation (see Figg. 87-90).

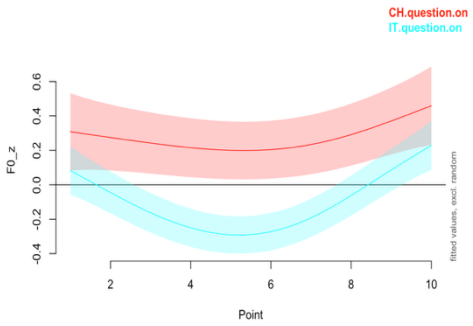


Figure 87 Question on-focus production by Language

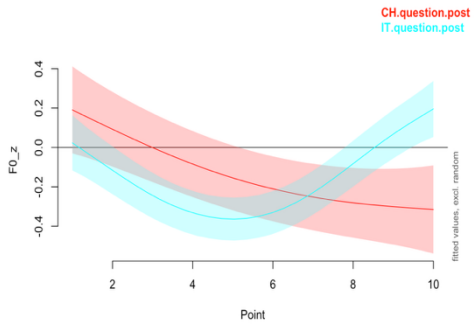


Figure 88 Question post-focus production by Language

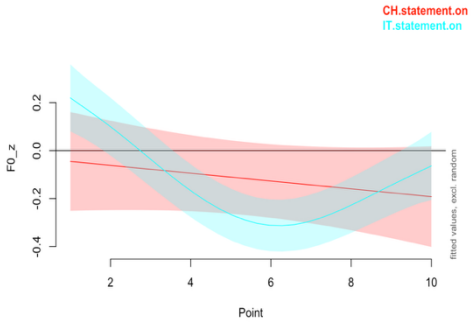


Figure 89 Statement on-focus production by Language

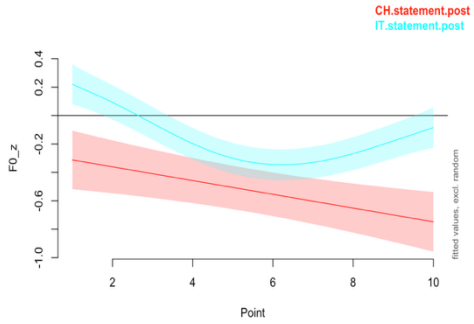


Figure 90 Statement post-focus production by Language

To facilitate direct comparisons of native vs. learner productions, a second model was fit with the three variables (Lang, S.Type, Focus) specified as separate factors and included the full interaction. EMMs were then extracted for each Lang.S.Type.Focus cell and pairwise contrasts were computed.

Table 38 Pairwise contrasts by Language across conditions

Condition	Estimate	SE	t-ratio	p-value	Interpretation
Question, on-focus	0.743	0.091	8.14	< .0001	Significant: CH > IT
Statement, on-focus	0.084	0.091	0.93	0.354	Not significant
Question, post-focus	0.281	0.091	3.08	0.0021	Significant: CH > IT
Statement, post-focus	-0.371	0.091	-4.07	< .0001	Significant: IT > CH

These results indicate IT showed consistently lower F0_z in question conditions, particularly under focus; whereas, in the post-focus position of statements, the pattern reversed: IT produced significantly higher F0_z than CH, suggesting a lack of PFC.

Using *ggplot2*, EMMs were visualized by plotting predicted F0_z values for each Lang.Focus combination within sentence types, with error bars representing 95% confidence intervals (see Fig. 91).

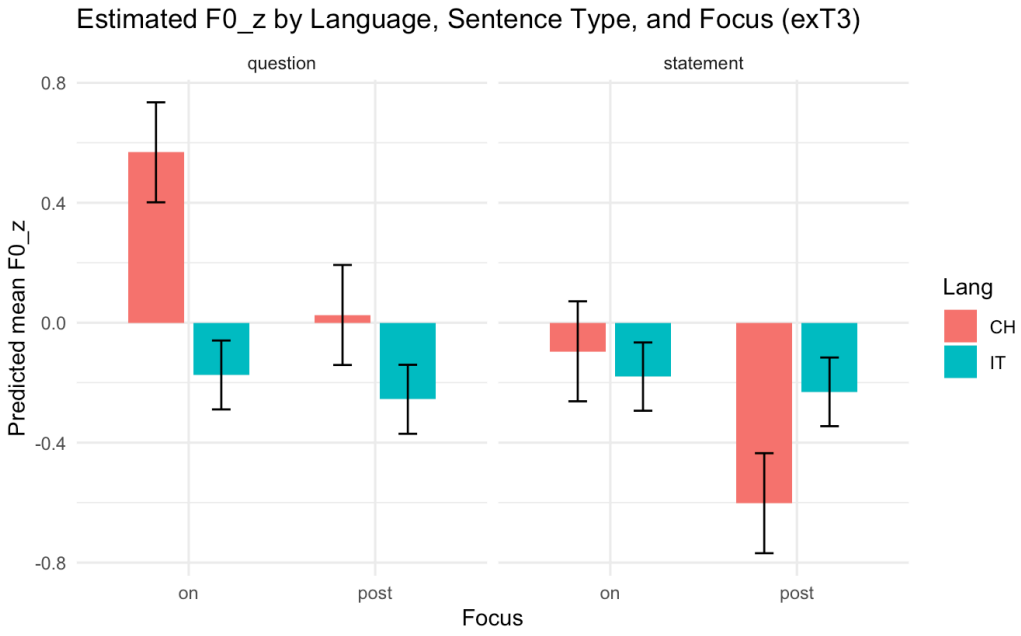


Figure 91 Estimated F0_z by Language, Sentence Type, and Focus (T3 excluded)

This exploratory analysis demonstrates that, within the experimental domain, Italian learners of Mandarin systematically diverge from native speakers in their tone production when encoding sentence-type intonation. Particularly in on-focus question contexts, learners generally undershoot pitch targets (cfr. Yang, 2016 on American learners of Mandarin), whereas in post-focus regions, learners generally fail to implement PFC, especially in statements.

In pursuit of a more specific and explanatorily robust model, the analysis was subsequently conducted within individual tone subsets, allowing for the identification of tone-dependent patterns in the data. For qualitative reference, Appendix I presents by-speaker contour plots for all tone subsets.

6.4 Tone 1 Subset Analysis

6.4.1 Interaction of Language, Sentence Type, and Focus

To examine the realization of T1 across different discourse configurations, we fit a GAMM to the subset of data corresponding to the second syllable of T1 disyllabic targets. The model included a three-way interaction between Language (CH vs. IT), Sentence Type (question vs. statement), and Focus (on-focus vs. post-focus), combined into a single composite factor (LSF). Time-varying pitch contours were modeled using smooths over the normalized time variable (Point) for each LSF level, with random smooths for Speaker and OtherTone to account for individual and contextual variability[23].

Parametric effects were significant for most contrasts relative to the CH.question.on baseline, with particularly large differences observed in the CH vs. IT contrast for statement.post (estimate = -0.95, $p < .0001$) and question.on (estimate = 0.77, $p < .0001$), indicating clear differences in how T1 is realized across sentence types in both on-focus and post-focus contexts by the two Language groups.

Smooth term results indicated that while several conditions yielded relatively flat trajectories (consistent with the level-contour nature of T1), significant time-varying F0 patterns were still present in a few conditions, notably in CH.question.on, which was characterized by a slight rising trend (see Fig. 92).

[23] The model converged successfully and explained approximately 34.9% of the deviance (adjusted R2=0.334), indicating a relatively strong fit given the limited pitch dynamics expected for T1.

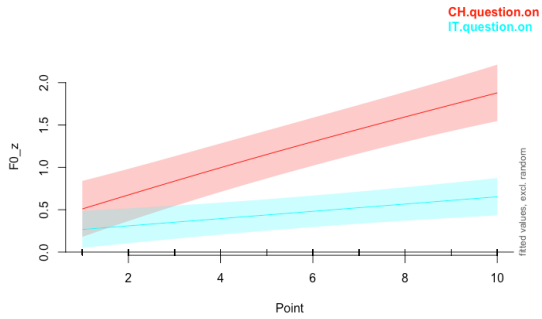


Figure 92 T1 question on-focus production by Language

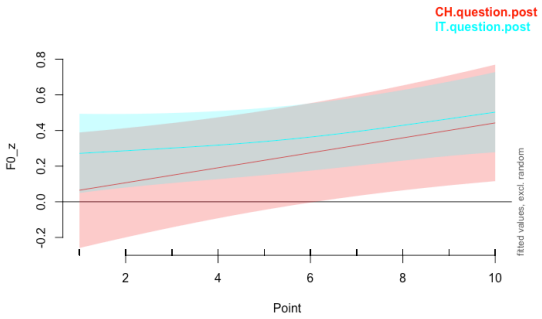


Figure 93 T1 question post-focus production by Language

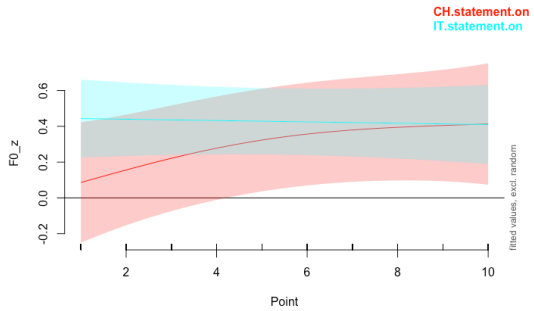


Figure 94 T1 statement on-focus production by Language

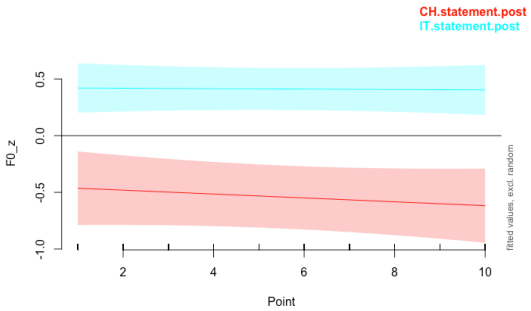


Figure 95 T1 statement post-focus production by Language

For greater interpretability, we re-fit the model with separate factors for Language, Sentence Type, and Focus, and conducted post hoc pairwise comparisons using EMMs. These comparisons confirmed that CH speakers produced significantly higher F0_z than IT speakers in question.on contexts ($p < .0001$); IT speakers produced higher F0_z than CH speakers in statement.post contexts ($p < .0001$); no significant differences were found for statement.on or question.post.

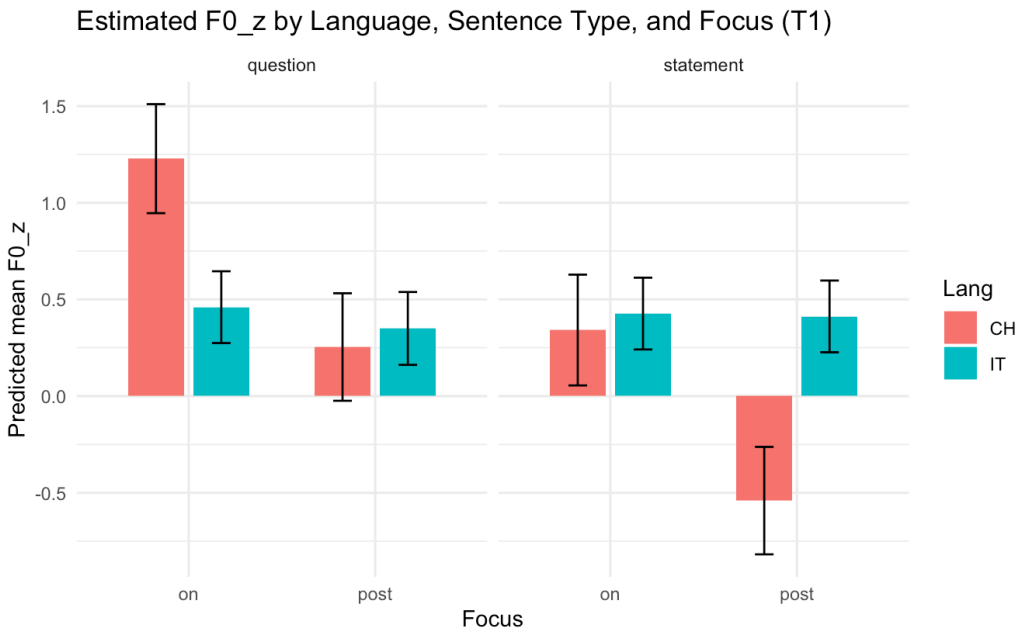


Figure 96 Estimated F0_z by Language, Sentence Type, and Focus (T1)

These results highlight the interaction between discourse structure and language background in the realization of what is typically assumed to be a stable, level pitch category. The observed differences suggest that Italian learners may not yet fully stabilize T1 in discourse-governed prosodic environments, especially in sentence-final positions or under focus. A more detailed examination is provided in the following section, where other curve parameters are also taken into account.

6.4.2 Analysis on Curve Parameters for Tone 1

6.4.2.1 F0 slope

To further characterize pitch dynamics in syllable-final T1 production, we analyzed the F0 slope (F0_slope) as an additional time-derived metric. For each utterance token, a linear regression was computed over these 10 time-normalized F0 values to extract the F0_slope, operationalized as the beta coefficient of the regression line in z-units per time unit. The fixed effects structure included a full factorial model with three predictors: Lang (CH vs. IT), S.Type (question vs. statement), and Focus (on-focus vs. post-focus). Two random intercepts were included: Speaker, to account for individual differences in pitch range and production strategy; OtherTone, representing the co-occurring tone in the first syllable of the disyllabic phrase.

A likelihood ratio test revealed that both random intercepts significantly contributed to model fit. Satterthwaite's approximation was used to assess fixed effects. Results showed that

main effects of Lang (F = 6.19, p = 0.0162), S.Type (F = 36.04, p < .0001), and Focus (F = 17.10, p < .0001) were all statistically significant. Two interactions reached significance: Lang.S.Type (F = 4.56, p = 0.033), and Lang.Focus (F = 11.71, p < .001), whereas the three-way interaction was non-significant (p = 0.378), as was the S.Type.Focus interaction (p = 0.077). Backward selection retained a reduced model including only the significant interactions (Lang.S.Type and Lang.Focus).

The main effect of Lang indicates that Italian learners produce overall shallower F0 slopes (more compressed or downward-trending contours) compared to native speakers. The main effect of S.Type suggests that statements are produced with flatter or more falling trajectories than questions, consistent with a global intonational effect. The main effect of Focus reflects the well-established PFC in Mandarin. The Lang.S.Type interaction reveals that the difference between questions and statements is less pronounced in IT learners, indicating that they may have difficulty modulating slope based on sentence type. The Lang.Focus interaction demonstrated that IT learners exhibit less PFC (i.e., higher slope values in post-focus contexts) relative to native speakers, suggesting a potential prosodic marking deficit or reduced control over post-focal pitch range.

EMMs for F0_slope by Language, across Focus (on vs. post) and S.Type (question vs. statement) are shown below:

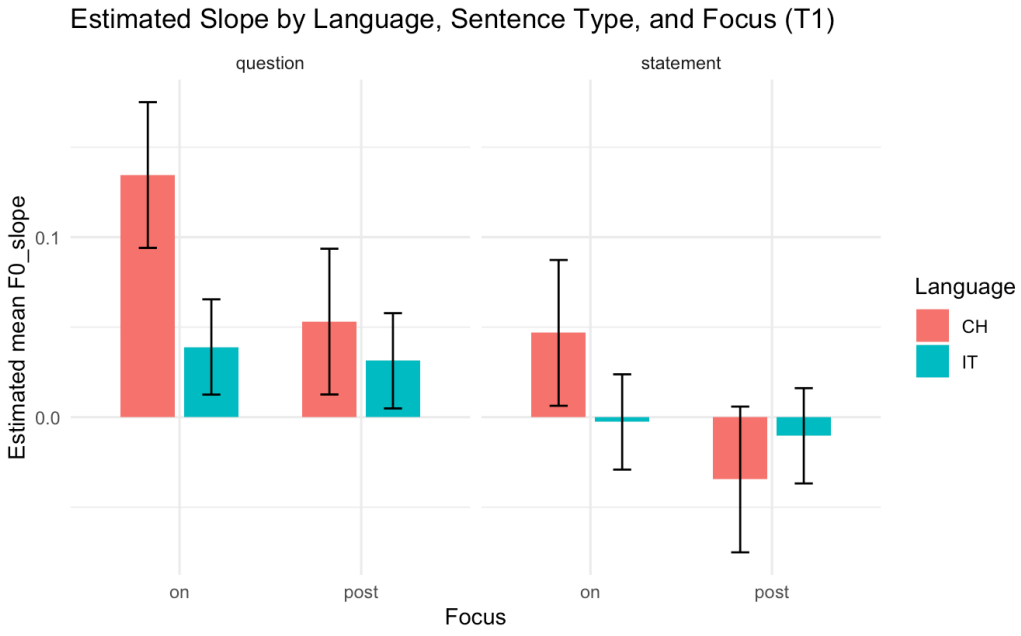


Figure 97 Estimated Slope by Language, Sentence Type, and Focus (T1)

These results demonstrate that IT differ markedly from CH in the dynamic shaping of their F0_slope, particularly in on-focus contexts (questions and statements) and in post-focus statements. The IT group's less evident contour modulation in questions and their failure to lower pitch in post-focus statements point to persistent L2-L1 differences in Mandarin intonation.

6.4.2.2 F0 max

To further characterize the pitch realization of T1 in sentence-final syllables, we analyzed the maximum fundamental frequency (F0_max), defined as the highest normalized F0 value across the 10 time-normalized points per utterance. This measure captures the upper bound of pitch realization and can be indicative of speakers' ability to achieve high tone targets.

We modeled F0_max as a function of Lang, S.Type and Focus. Random intercepts were specified for Speaker and OtherTone. A random effects structure evaluation revealed that both intercepts contributed significantly to model fit. Satterthwaite's approximation showed that main effects of Lang ($F = 3.70$, $p = 0.0582$), S.Type ($F = 87.65$, $p < .0001$), and Focus ($F = 36.45$, $p < .0001$) were all statistically significant. Two interactions reached significance: Lang.S.Type ($F = 86.93$, $p = 0.0001$), and Lang.Focus ($F = 35.11$, $p < .0001$), whereas the three-way interaction was non-significant, as was the S.Type.Focus interaction.

Model simplification eliminated the non-significant two-way and three-way interactions and converged successfully (REML = 3607.1), explaining substantial variance in F0_max.

The main effect of Lang indicates that IT produced significantly lower maximum F0 values compared to CH overall.

Statements were associated with a lower F0_max than questions, as captured by the strong main effect of S.Type, consistent with Mandarin's sentence-type prosody where questions typically involve greater pitch expansion (see § 2.6.1). Post-focus syllables exhibited lower F0_max than on-focus counterparts, reflecting PFC effect.

The Lang.S.Type interaction reveals that the difference in F0_max between questions and statements is reversed or significantly reduced for IT. While CH raised F0_max more in questions, IT showed a diminished or even opposite pattern.

Similarly, the Lang.Focus interaction indicates that PFC was less pronounced among L2 learners, suggesting potential difficulty in realizing the fine-grained prosodic adjustments required.

These findings underscore that learners differ markedly from native speakers in realizing pitch maxima, especially when modulating F0 according to sentential structure or discourse-level prominence.

Estimated F0_max values were visualized across Focus conditions (on vs. post) and Sentence Types (question vs. statement), separately for CH and IT in Fig. 98 below:

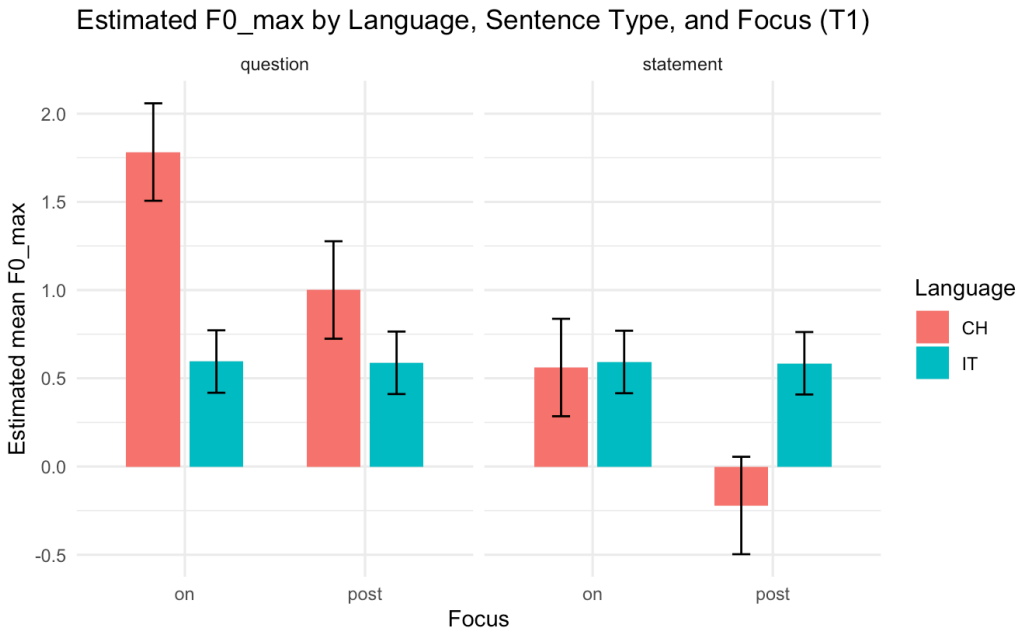


Figure 98 Estimated F0_max by Language, Sentence Type, and Focus (T1)

6.4.2.3 F0 min

The minimum fundamental frequency (F0_min), defined as the lowest normalized F0 value across the 10 time-normalized points per utterance, was calculated in the dataset for the analysis. This measure reflects the lower bound of pitch realization and is particularly relevant for capturing patterns of PFC and overall pitch floor, which can vary across languages and focus contexts. We modeled F0_min as a function of Lang, S.Type, and Focus. Random intercepts were specified for Speaker and OtherTone. Random structure evaluation proved that both random intercepts significantly improved model fit. Satterthwaite's approximation indicated that all three main effects were significant: Lang ($F = 13.35$, $p = .0005$), S.Type ($F = 9.00$, $p = .0027$) and Focus: $F = 50.55$, $p < .0001$; and two interaction terms also reached significance: Lang.S.Type ($F = 23.61$, $p < .0001$) and Lang.Focus ($F = 38.77$, $p < .0001$). In contrast, the S.Type.Focus interaction and the three-way interaction Lang.S.Type.Focus were non-significant and were removed in model simplification. The final model converged successfully (REML = 3335.6), retaining the main effects and the two significant two-way interactions.

The main effect of Lang showed that IT produced lower pitch minima overall, although this difference was marginally non-significant in the final model (p = .123). Statements were associated with lower F0_min than questions, consistent with a general downward pitch trend in non-interrogative contexts. Similarly, post-focus syllables exhibited lower pitch minima compared to on-focus tokens, indicating expected PFC.

Critically, the Lang.S.Type interaction demonstrated that the pitch-lowering effect observed in statements was significantly attenuated or even reversed in Italian learners, suggesting potential difficulties in suppressing pitch appropriately in non-interrogative utterances. Likewise, the Lang.Focus interaction indicated that PFC was less pronounced in the IT group compared to CH.

These findings highlight that IT diverge from CH not only in achieving pitch peaks (F0_max) but also in implementing pitch floor modulation.

Estimated F0_min values were visualized across focus conditions (on vs. post) and sentence types (question vs. statement), separately for CH and IT in Fig. 99 below:

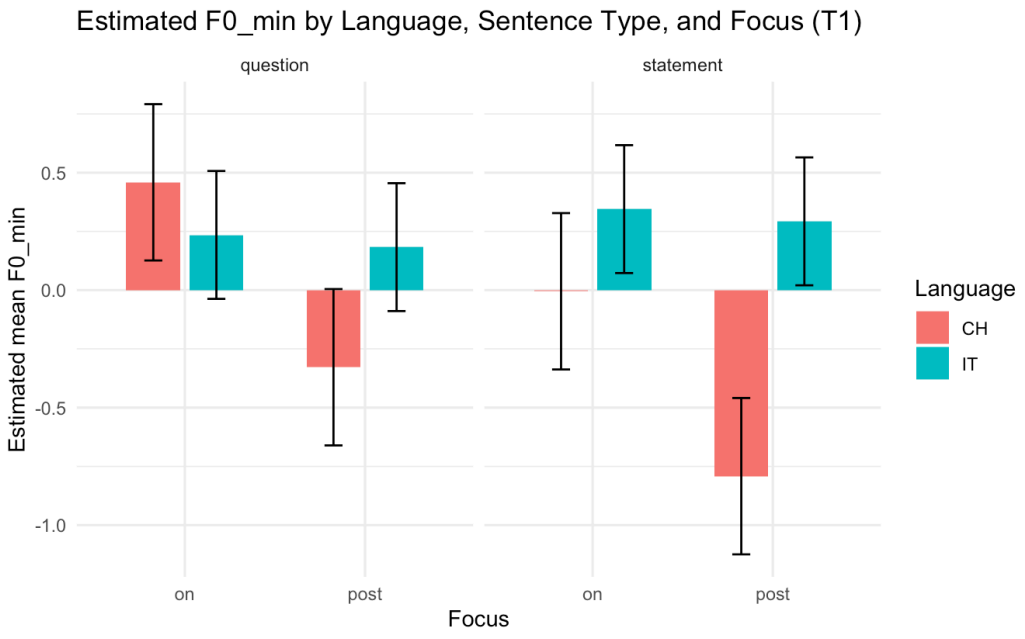


Figure 99 Estimated F0_min by Language, Sentence Type, and Focus (T1)

6.4.2.4 F0 range

To provide a more nuanced understanding of pitch modulation in sentence-final T1, we examined the pitch range (F0_range) – defined as the difference between maximum and minimum normalized F0 values across the 10 time-normalized sampling points for each utterance.

A GLMM was fit to the data, with Language, Sentence Type, and Focus as fixed effects, and random intercepts for Speaker and OtherTone. Both random effects significantly contributed to the model fit (Speaker: $\chi^2(1) = 48.09$, $p < .0001$; OtherTone: $\chi^2(1) = 9.06$, $p = .0026$), confirming the presence of speaker- and tonal-contextual variability in pitch range realization.

The fixed effects analysis revealed highly significant main effects of Language ($F(1, 73.48) = 40.88$, $p < .0001$), Sentence Type ($F(1, 1444.80) = 45.81$, $p < .0001$), and a marginally non-significant effect of Focus ($F = 0.15$, $p = .697$).

Critically, a significant three-way interaction emerged between Language.Sentence Type.Focus ($F(1, 1444.8) = 3.98$, $p = .0461$), justifying the retention of all interaction terms in the final model. The model converged successfully (REML = 3556.7), explaining substantial variance in F0_range and validating the complexity of prosodic interactions across speaker groups and discourse contexts.

The main effect of Language demonstrates that Italian learners produced significantly reduced pitch spans overall compared to native speakers, consistent with a more constrained prosodic realization. Similarly, statements were associated with narrower pitch range than questions, reflecting Mandarin's interrogative pitch expansion pattern. While focus did not independently affect F0_range in T1 targets, its role became evident in interaction with language background and sentence type.

The Language.Sentence Type interaction revealed that CH exhibited a wider pitch range in questions relative to statements, while IT produced a diminished or even reversed pattern, suggesting reduced sensitivity to sentence-level prosodic cues. This supports the view that sentence type-specific pitch adjustments remain challenging for Italian learners.

The three-way interaction showed that PFC, which should manifest as a narrower pitch range, was less consistent in Italian productions, particularly in statements. This finding aligns with broader evidence of difficulty in implementing prosodic de-accentuation and supports the idea that focus-related pitch modulation (both enhancement and suppression) poses persistent challenges in L2 acquisition.

In sum, the results underscore the interdependence of pitch range with language experience, sentence type, and information structure, and they reveal that while native speakers dynamically modulate pitch span according to both sentence type and focus, learners display a more limited prosodic flexibility, particularly in contexts requiring PFC or sentence-type driven pitch expansion.

Visualizations of EMMs confirmed these patterns, showing attenuated prosodic contrasts in the L2 group across Focus and Sentence Type conditions:

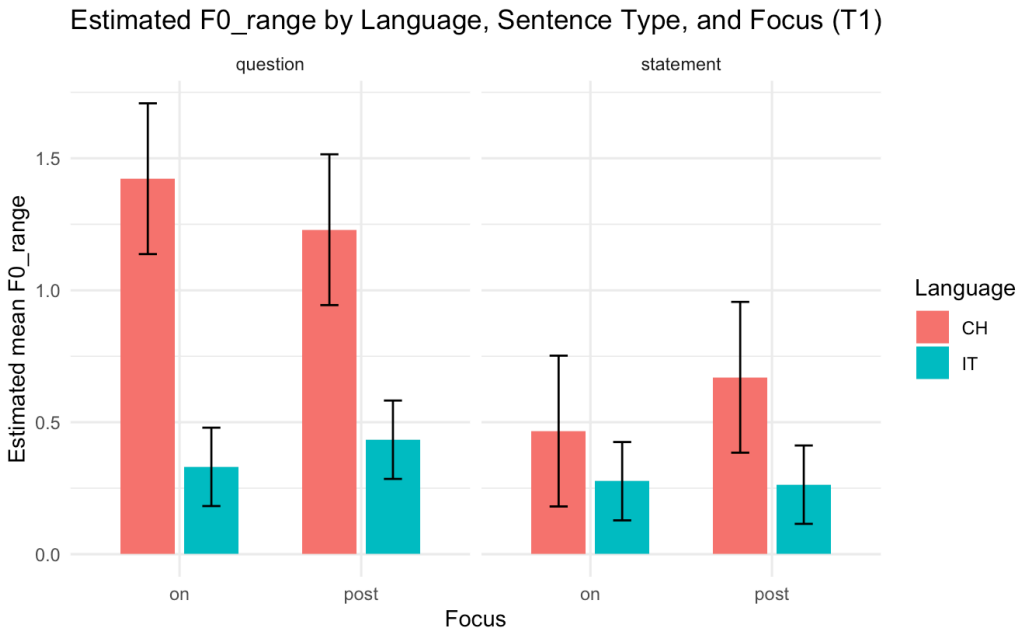


Figure 100 Estimated F0_range by Language, Sentence Type, and Focus (T1)

6.4.3 Italian learner subset

To further examine how subgroups of Italian learners, classified by Proficiency, Musicality, and academic Grade, modulate their pitch contours for T1 in response to sentence-level conditions, the analysis was restricted to the Italian learner dataset.

6.4.3.1 Comparing learner-factor models (Proficiency, Musicality, Grade)

To explore which learner-specific factors best account for variation in T1 production among Italian L2 speakers, we conducted a series of model comparisons using maximum likelihood (ML) estimation. The base model included only sentence-level factors – Sentence Type and Focus (SF model) – as predictors of normalized pitch (F0_z). This model served as a benchmark against which models incorporating Proficiency, Musicality, and Grade were tested for their additional explanatory power. Each enriched model used a composite predictor:

PSF: Proficiency.Sentence Type.Focus

MSF: Musicality.Sentence Type.Focus

GSF: Grade.Sentence Type.Focus

All models included identical random smooths for Speaker and OtherTone, and fixed smooths for the F0 contour over normalized time (Point), by interaction level. Each enriched model was compared to the base SF model using likelihood ratio tests and differences in AIC.

As reported in Tab. 39, the model with Proficiency yielded a highly significant improvement in fit over the base model ($\chi^2 = 24.73, df = 12, p < .001$), and reduced AIC by 38.5 points. This suggests that learners' pitch trajectories are strongly modulated by their proficiency level.

In contrast, the model incorporating Musicality did not significantly improve fit (p = .467) and provided no AIC benefit (ΔAIC = -0.95), indicating that musical skill was not a strong predictor of T1 F0 contours in this experimental context.

The Grade model showed a moderate but significant improvement over the base model (χ^2 = 23.71, df = 24, p = .003), with a 16.3-point reduction in AIC. This suggests that formal academic progression does play some role in shaping tone production, although the effect is weaker than that of proficiency.

Table 39 Proficiency, Musicality and Grade model comparison with baseline

Comparison	Compared edf	ΔLR χ^2	df	p-value	Sig.
SF vs PSF	28	24.729	12	1.74×10^{-6}	***
SF vs MSF	28	5.870	12	0.467	
SF vs GSF	40	23.714	24	0.003	**

To directly compare the top-performing models, we extracted their AIC values in Tab. 40 below:

Table 40 Comparison of mPSF, mGSF, and baseline models

Model	edf	AIC
mPSF_ml	155.4621	15 129.98
mGSF_ml	163.3426	15 152.14
m_SF_ml	150.2061	15 168.47

The lowest AIC was observed for the Proficiency model (mPSFT1_ml), confirming it as the best-fitting model among all tested. The Grade model (mGSF_ml) ranked second, followed by the baseline SF model.

6.4.3.2 Interaction of Proficiency, Sentence Type and Focus

To further investigate the influence of individual language proficiency on the production of Mandarin T1 in disyllabic words, we refit the PSF model with method = "fREML"[24].

The fixed effect intercept corresponded to the reference level (PSFHigh.question.on), with an estimated F0_z of 0.55, significantly above zero ($t = 3.62$, $p < .001$), suggesting that high proficiency learners produced a high and stable pitch in question-on-focus contexts, consistent with canonical T1 targets.

Two significant contrasts emerged. In statement.on-focus sentences, high proficiency learners produced significantly lower F0 compared to the reference, with a mean difference of -0.104 ($p = .046$). In statement.post-focus sentences, the high-proficiency group also exhibited a significant lowering of pitch (-0.120, $p = .022$). No significant effects were found for low proficiency learners in any context, although the Low.question.post contrast approached significance ($p = .106$), hinting at possible emerging patterns. These results indicate that high-proficiency learners are more sensitive to sentence-level prosodic conditions, adjusting their pitch targets for T1 in a context-dependent manner, especially under more neutral conditions (i.e., post-focus or non-question).

The contour shapes – captured by smooth terms for Point within each PSF level – provided further evidence of proficiency-related modulation in pitch trajectories.

Smooths for the question condition revealed that low-proficiency learners exhibited a significant curvilinear pattern (edf = 2.40, F = 8.21, p < .001), indicating noticeable pitch movement within the tone contour. In contrast, high-proficiency learners produced an almost flat and stable trajectory (edf = 1.00, p = .22), closely approximating the monotonic high-level pitch characteristic of the T1 citation form, though still deviating from native-like production in this experimental condition (see Fig. 101).

[24] Model diagnostics indicated stable convergence and acceptable basis dimension choices (k-index ≈ 0.98 for all terms), suggesting that the spline bases were adequate to capture the contour shape without overfitting.

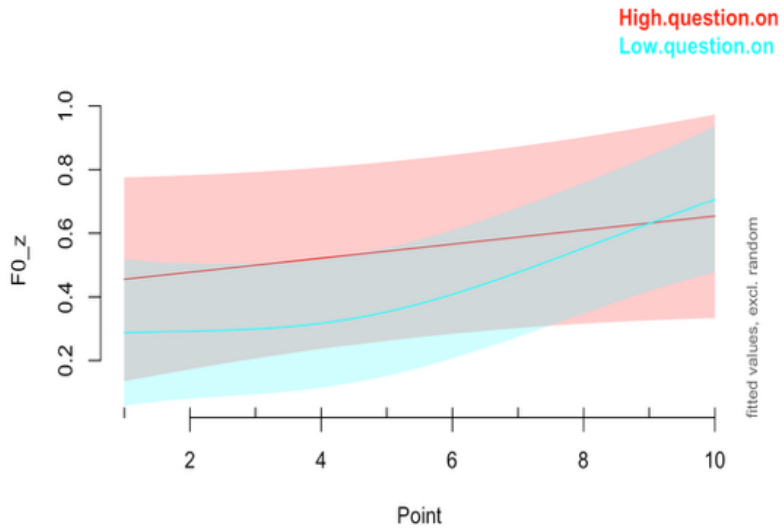


Figure 101 Tone 1 question on-focus production by Proficiency

Smooths for question.post prove that both proficiency groups displayed curved trajectories, but the effect was slightly more pronounced for low proficiency (edf = 3.24, F = 8.55, p < .001) than for high proficiency (edf = 3.17, F = 3.29, p = .010). Low proficiency learners' predicted curve also exhibit a slight rise trend starting on the second portion of the curve, while a falling trend characterized the second portion of the high proficiency learners' curve (see Fig. 102).

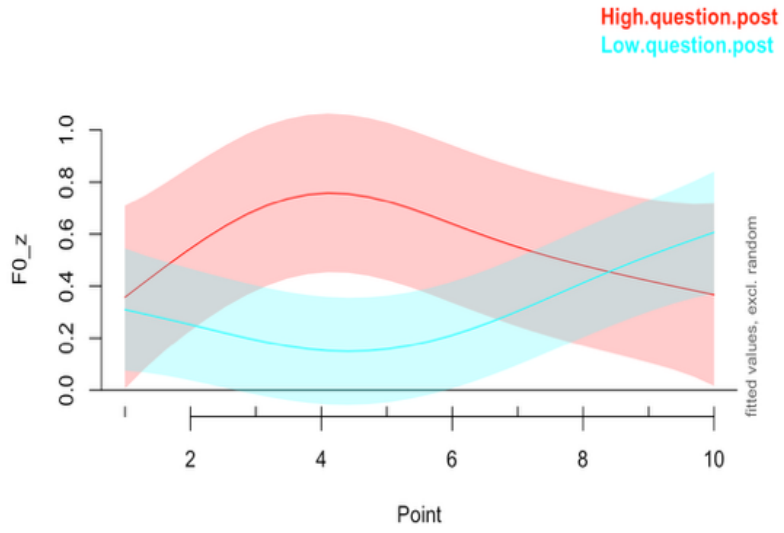


Figure 102 T1 question post-focus production by Proficiency

Smooths for statement.on and statement.post prove that, across these two conditions, only low proficiency learners in the statement.post condition showed a slight trend toward curvature (edf = 2.49, p = .08), while all other curves remained largely flat (see Figg. 103-104).

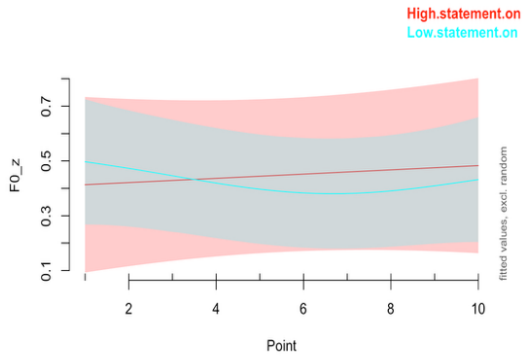


Figure 103 T1 statement on-focus production by Proficiency

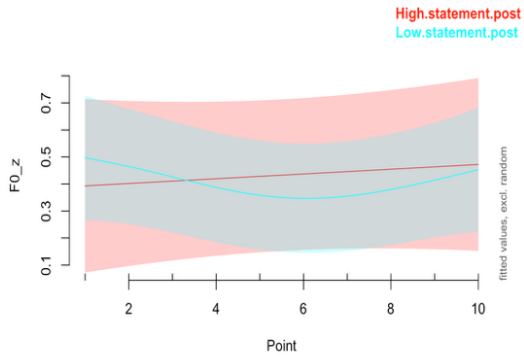


Figure 104 T1 statement post-focus production by Proficiency

The results of the GAMM analysis indicate that linguistic proficiency exerts a substantial influence on the phonetic realization of Mandarin T1 among Italian learners, with effects being particularly pronounced in interrogative contexts. Learners with higher proficiency demonstrate a greater ability to preserve tonal contours under context-sensitive conditions, producing realizations that more closely approximate tonal targets in their citation form. Nonetheless, systematic divergences from native speaker patterns persist, most notably in the form of a limited rising tendency in question-on conditions and a falling-like tendency in question-post conditions.

By contrast, learners with lower proficiency exhibit increased pitch variability and reduced consistency in aligning tonal production with prosodic structure. These deviations are especially salient in post-focus environments, where native speakers typically employ deaccentuation strategies. These patterns suggest that less proficient learners may rely, at least in part, on prosodic transfer from their L1, whereas higher-proficiency learners are better able to preserve the citation form of the target tone across experimental conditions.

6.4.3.3 Interaction of Grade, Sentence Type and Focus

To evaluate the influence of academic progression on the production of Mandarin T1 among second-language learners, we refitted the GSF model with method = "fREML"[25].

Inspection of the parametric coefficients revealed several noteworthy trends. The reference level was set to second-year undergraduate learners in the question.on-focus condition (GSF = BA2.question.on). Relative to this baseline, first-year Master's students (MA1) produced significantly higher F0_z values in both the question-on and statement-on contexts. Specifically, MA1 learners exhibited a 0.34 increase in F0_z in the question-on condition (p = .0236), and a 0.39 increase in the statement-on condition (p = .0094). A marginal increase was also observed for MA1 in the statement-post condition (p = .0635), suggesting a consistent elevation of pitch across prosodic contexts. In contrast, no significant differences were found between BA2 and BA3 students, indicating that tonal refinement may not progress linearly across the undergraduate curriculum but rather may undergo more substantial development at the postgraduate level.

Turning to the smooth terms, the analysis of F0 contours across time revealed important differences in the shape and stability of pitch trajectories. In the question-on context, the smooth term for BA3 was statistically significant and displayed moderate curvature (*edf* = 1.44, p < .0001), suggesting that these learners may struggle to maintain a flat, target-like pitch throughout the syllable. Interestingly, although MA1 learners exhibited higher overall pitch, their smooth was effectively linear (*edf* = 1.00), indicative of a more controlled and consistent T1 production. BA2 learners showed a modest degree of curvature (*edf* = 1.88), though this did not reach statistical significance (Fig. 105).

[25] Model diagnostics confirmed a good fit, with no convergence issues or signs of oversmoothing. The basis dimension check yielded a k-index close to 0.99 for all smooths, indicating that the model's flexibility was sufficient to capture the temporal dynamics of F0 trajectories. The adjusted R-squared value was 0.27, with approximately 28.7% of deviance explained.

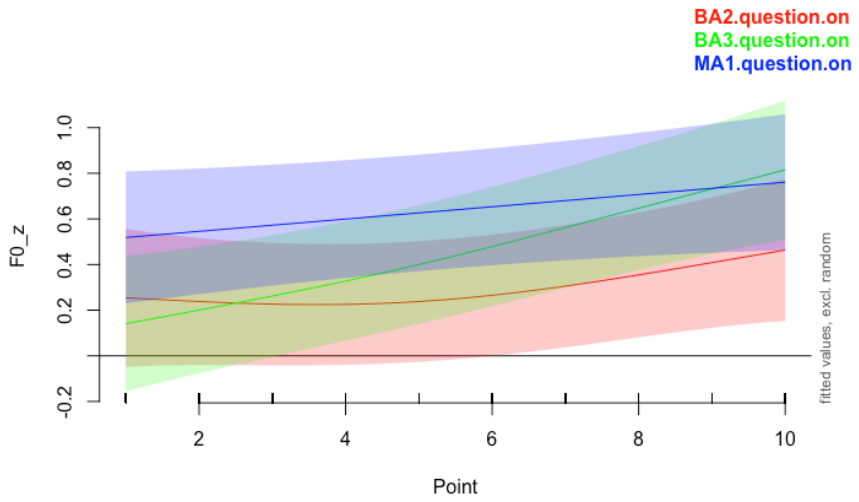


Figure 105 T1 question on-focus production by Grade

In statement-on conditions, none of the smooths reached significance. The absence of significant pitch modulation in this context aligns with the canonical realization of T1 as a high-level tone, particularly when not influenced by interrogative intonation (see Fig. 106).

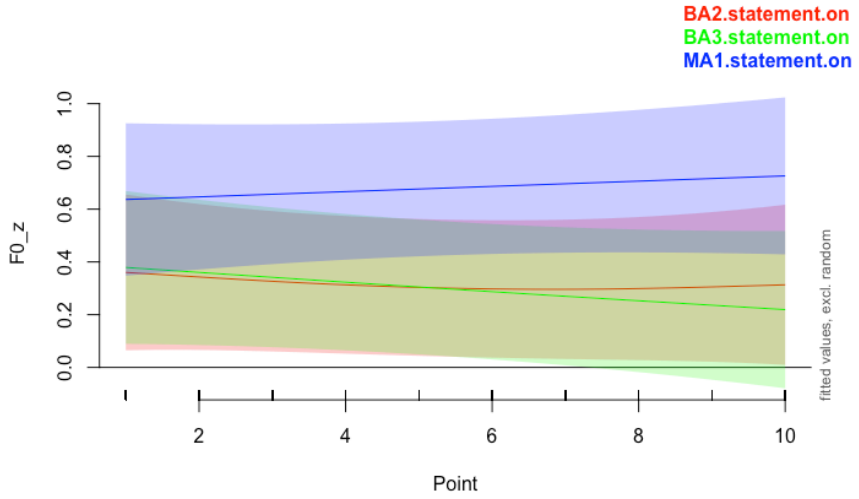


Figure 106 T1 statement on-focus production by Grade

By contrast, the question-post condition produced more differentiated results. Both BA2 and BA3 learners exhibited significant curvature, suggesting increased contour instability in this post-focal, sentence-final environment. For BA2, the smooth had an estimated degrees of freedom of 2.50 and reached significance (p = .036), while for BA3, the effect was again significant, despite a linear shape (edf = 1.00, F = 7.72, p = .0055). These results imply that, in the absence of strong focal marking, learners at the BA level may exhibit a degradation of tonal

precision, potentially due to deaccenting or sentence-final intonation override. The MA1 group, although not exhibiting significant contour variation, maintained higher overall pitch in this context, suggesting an improved ability to preserve tonal targets across conditions (see Fig. 107).

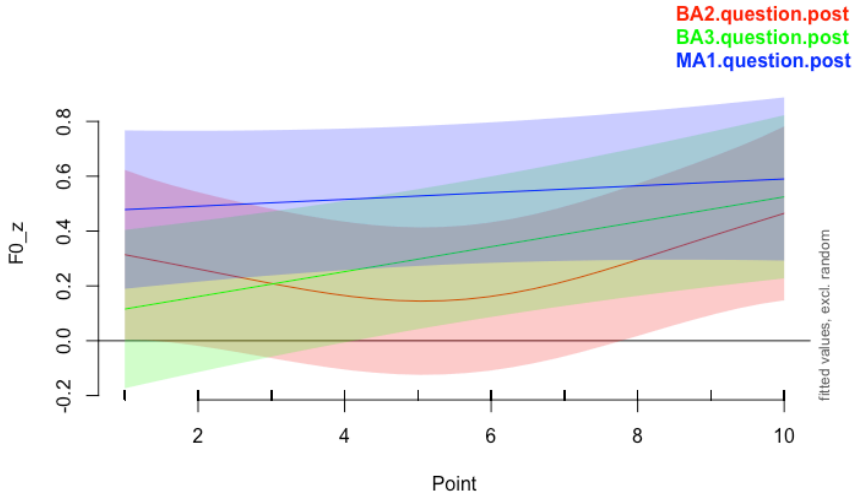


Figure 107 T1 question post-focus production by Grade

Finally, in the statement-post condition, no group showed significant temporal variation in pitch contour. While parametric estimates for MA1 learners were marginally higher than the BA groups, the smooth terms did not differ significantly, suggesting limited modulation of pitch over time in this neutral post-focal environment (see Fig. 108).

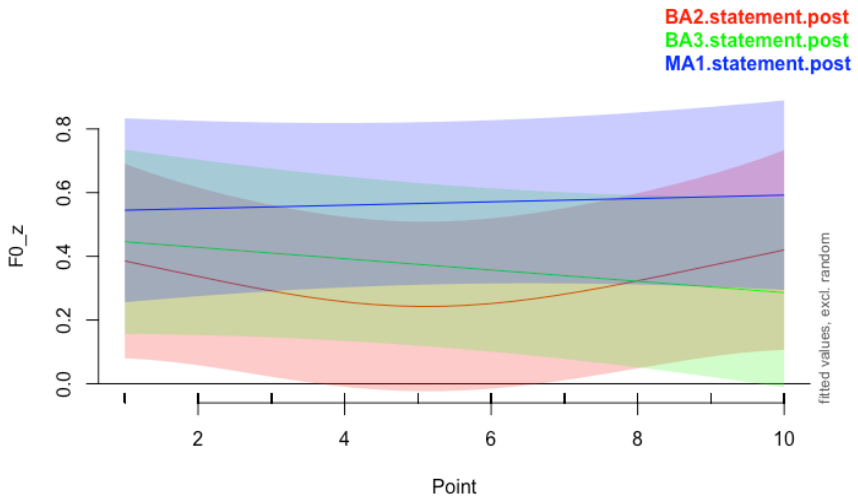


Figure 108 T1 statement post-focus production by Grade

Taken together, these findings on Grade model suggest that academic progression has a measurable, though selective, influence on the production of Mandarin T1 in Italian learners. While undergraduate learners (BA2 and BA3) exhibit relatively unstable tone production under prosodically demanding conditions such as sentence-final or post-focal contexts, postgraduate learners (MA1) demonstrate a more target-like realization of T1 citation form, characterized by higher pitch values and flatter, more stable contours.

6.4.4 Interim summary on T1

The results of the T1 subset analysis highlight systematic differences between native Mandarin speakers and Italian learners in the prosodic realization of T1 in sentence-final position. Although T1 is canonically level, the GAMM and subsequent GLMM analyses revealed that both groups modulate its pitch trajectory in response to discourse-level factors such as sentence type and focus, albeit in different ways. Native speakers exhibited consistent and context-sensitive adjustments, including elevated F0 in questions, PFC, and wider pitch range in interrogative contexts – reflecting a dynamic and integrated use of pitch for both lexical tone and sentence type. In contrast, Italian learners displayed shallower pitch contours, lower F0 maxima, and narrower pitch ranges overall, with reduced differentiation between focus conditions and sentence types. Notably, interactions involving sentence type and focus revealed that learners often underutilized prosodic cues such as PFC or interrogative pitch expansion, suggesting difficulties in coordinating tonal and intonational cues. These results point to persistent L1-L2 mismatches in prosodic encoding, which may stem from limited exposure to Mandarin speech, transfer from Italian intonational patterns, and/or reduced phonological awareness of sentence-level tonal modulation.

The T1 analysis within the Italian learner subset reveals that individual learner factors – particularly language proficiency and academic progression – exert a significant influence on the realization of Mandarin's high-level tone in sentence-final position.

The GAMM analysis confirms that Proficiency modulates the realization of Mandarin T1 among Italian learners, particularly in relation to sentence type and focus condition. High-proficiency learners displayed context-sensitive adjustments in F0: they produced significantly lower pitch in both statement-on and statement-post conditions relative to question-on, while maintaining a stable high pitch in question-on focus contexts, consistent with canonical T1 targets. By contrast, low-proficiency learners showed no significant effects across conditions, although a near-significant trend emerged in the question-post condition.

The analysis of pitch trajectories further highlighted Proficiency effects. In question-on focus conditions, low-proficiency learners exhibited curvilinear F0 contours, indicating pitch movement within the tone, whereas high-proficiency learners produced flat trajectories that closely approximated the expected high-level T1 citation form.

In question-post contexts, both groups showed contour curvature, but the rising trend in low-proficiency learners contrasted with the slight falling tendency observed in the high-proficiency group. Statement conditions yielded largely flat contours, with only a weak tendency toward curvature in low-proficiency learners under post-focus. Overall, the findings suggest that high-proficiency learners are more adept at implementing context-dependent tonal adjustments, thereby approximating native-like targets, although certain divergences remain in interrogative contexts. Low-proficiency learners, in contrast, revealed greater pitch variability and appear less attuned to prosodic conditioning, possibly reflecting increased reliance on L1 prosodic strategies.

Academic grade level highlighted a secondary influence: MA1 learners consistently produced higher and flatter contours across contexts, suggesting lexical tone stabilization. However, despite producing citation-like T1 shapes, these learners showed limited modulation across pragmatic contexts, indicating that lexical tone mastery may not guarantee full prosodic integration. This finding highlights a developmental dissociation between tonal accuracy and sentence-level intonation control. Ultimately, these results underscore that successful L2 tone acquisition in Mandarin is not solely a matter of phonemic precision but depends on learners' ability to integrate tonal and intonational cues in a context-sensitive manner – a skill that may require explicit instruction in prosody and discourse-level pragmatics.

This difference supports the hypothesis that while MA1 learners prioritize segmental accuracy (i.e., producing citation-like tones), more proficient learners are better equipped to navigate the interface between lexical tone and sentence-level prosody. This distinction carries important implications for second language tone acquisition. It suggests that prosodic nuance may develop on a separate timeline from lexical tone accuracy, potentially requiring more targeted training in intonation and discourse pragmatics. Further perceptual research is needed to evaluate whether native listeners perceive MA1 tonal productions as pragmatically informative or merely lexically accurate. Such studies could provide crucial evidence as to whether learners at this level are able to communicate meaning beyond the word level, particularly in tonal languages where pitch serves both phonemic and prosodic functions.

6.5 Tone 2 Subset Analysis

6.5.1 Interaction of Language, Sentence Type, and Focus

To examine how T2 is realized across conditions and between speaker groups, we fitted a GAMM incorporating a three-way interaction between Language background (CH vs. IT), Sentence Type (question vs. statement), and Focus position (on-focus vs. post-focus). The model was specified using the interaction factor LSF (= Lang.S.Type.Focus) as the main predictor, with by-smooths over *Point* for each level of LSF, along with random smooths over *Speaker* and *OtherTone* to account for by-speaker variation and tonal context effects[26].

The parametric coefficients revealed significant baseline differences in F0_z across language and pragmatic contexts. Notably, the CH.statement.on condition exhibited significantly lower F0_z values compared to the reference level (CH.question.on), with an estimate of -0.32 ($p < .0001$). Similarly, CH.statement.post was characterized by an even more pronounced drop, with an estimate of -0.47 ($p < .0001$). In contrast, none of the IT conditions differed significantly from the reference, although IT.question.on approached marginal significance (estimate = +0.21, $p = .062$). These results suggest that CH modulate T2 pitch realizations more robustly across sentence types and focus conditions than IT, whose values remained closer to baseline across contexts.

Smooth terms revealed highly significant time-varying patterns in most LSF conditions. All IT conditions involving on-focus or post-focus positioning in both sentence types exhibited strong nonlinear effects (all $p < .0001$), indicating dynamic pitch movement over time. In the CH group, significant time-varying pitch trajectories were observed in the CH.question.on, CH.question.post, and – albeit more weakly – CH.statement.on conditions. In contrast, the CH.statement.post condition, despite exhibiting a significantly lower parametric intercept, did not yield a statistically reliable smooth term ($p = .14$). This absence of temporal variation in the F0 contour suggests a flattened or compressed pitch trajectory, likely reflecting tonal undershoot in this post-focal, sentence-final context. Such underspecification aligns with prior observations of PFC in Mandarin, where tonal targets may be acoustically attenuated in deaccented prosodic environments (Xu, 1999; Chen, 2010).

[26] The model showed adequate convergence and diagnostic performance. Basis dimension checking (k-index ≈ 0.99) suggested no undersmoothing across conditions. The model explained approximately 33% of the deviance (R2=0.31), indicating a moderate but meaningful degree of model fit given the complex interaction structure and inherent variability in prosodic data.

These patterns point to a robust tone-intonation interface in native Mandarin production, whereby tonal realization is dynamically adapted according to discourse structure. In contrast, Italian learners appear to maintain tone shapes more uniformly, possibly reflecting a greater reliance on lexical pitch targets than on prosodic modulation (see Figg. 109-112).

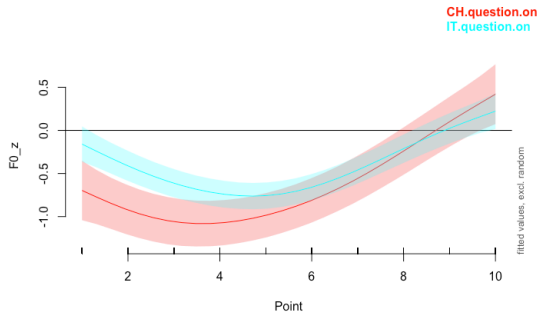


Figure 109 Tone 2 question on-focus production by Language

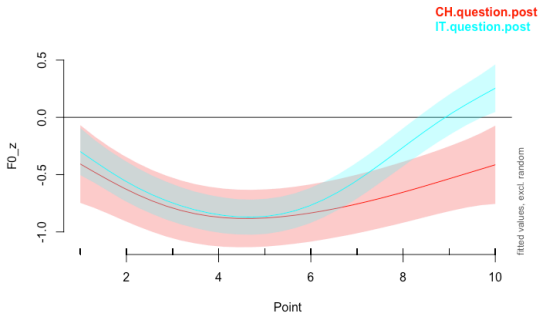


Figure 110 Tone 2 question post-focus production by Language

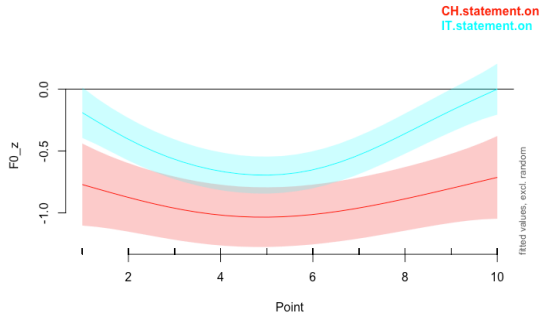


Figure 111 Tone 2 statement on-focus production by Language

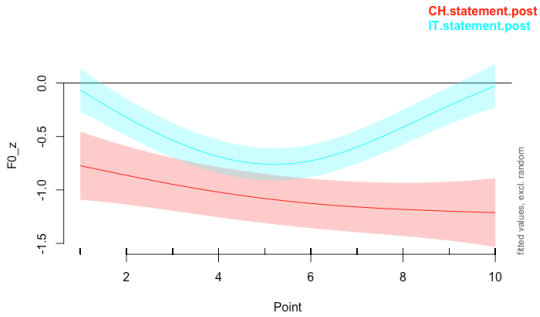


Figure 112 Tone 2 statement post-focus production by Language

To isolate differences between CH and IT speakers across all pragmatic contexts, we refitted the GAMM using three separate predictors – Language, Sentence Type, and Focus – and extracted EMMs. The Lang.S.Type.Focus interaction was interrogated through pairwise contrasts. The following significant differences emerged:

Table 41 Pairwise contrast across Language

Sentence Type	Focus Position	CH-IT Estimate	SE	p-value	Significance
Statement	On	-0.3449	0.132	0.0089	**
Statement	Post	-0.3480	0.128	0.0064	**
Question	On	-0.1847	0.139	0.1848	n.s.
Question	Post	-0.0299	0.136	0.8261	n.s.

These findings indicate that CH lower F0_z significantly more than IT in statement contexts, both under focus and post-focus positioning. The lack of significant effects in interrogative contexts indicates that learners' tone realizations converge more closely with native patterns, likely because the phonological rising contour aligns with the expected global intonational rise marking interrogativity in learners' L1 varieties.

A bar plot of predicted marginal means confirms these patterns (see Fig. 113): in both statement.on and statement.post conditions, CH speakers consistently exhibit lower F0_z than IT speakers. In contrast, in both question conditions, F0_z values between the two groups are much closer, with only slight differences.

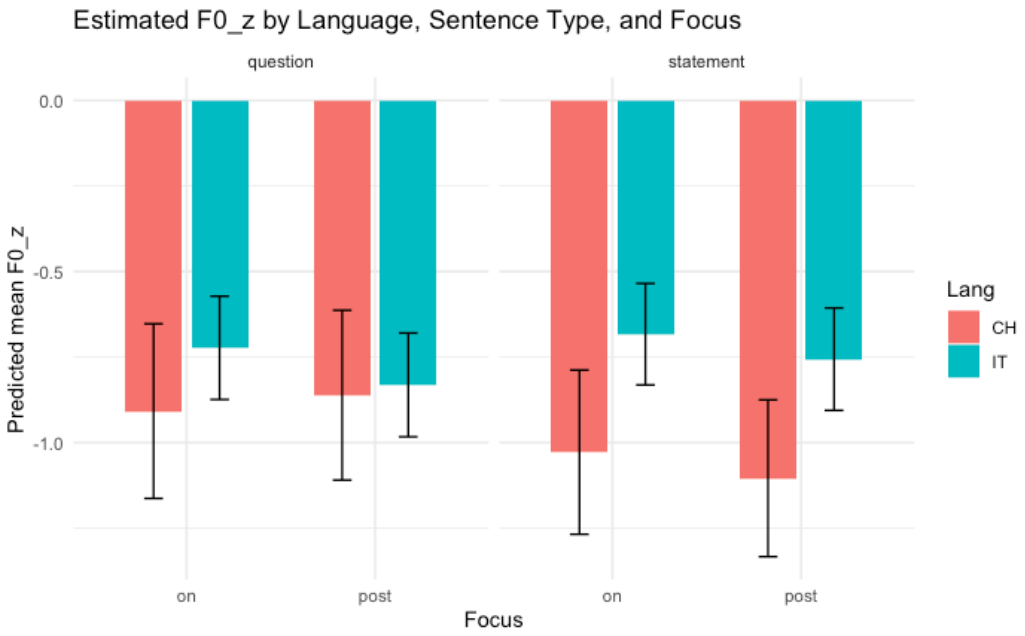


Figure 113 Estimated F0_z by Language, Sentence Type, and Focus

Together, these findings highlight the prosodic complexity involved in the production of rising tones within sentence-level contexts, and the challenges faced by Italian learners of Mandarin in coordinating lexical tone with intonational structure. While learners demonstrate a reasonable ability to produce T2 accurately in isolation or in interrogative on-focus contexts – conditions that align closely with the citation form – they appear less adept at modulating pitch in response to focus structure or declarative sentence type. This pattern supports the view that surface-level acquisition of tonal categories may outpace learners' internalization of their context-sensitive phonetic implementations.

Moreover, the robust nonlinear pitch excursions observed in learner productions – even in conditions where native speakers show flattened contours due to PFC or tonal undershoot – suggest a possible overapplication of canonical tone contours. This tendency may reflect an instructional bias toward segmental tone accuracy at the expense of prosodic economy and discourse pragmatics, especially in pedagogical contexts where intonation receives limited attention. These results invite further inquiry into whether learners systematically overgeneralize citation-form tone contours without fully integrating them into the intonational and pragmatic structure of connected speech.

6.5.2 Analysis on Curve Parameters for Tone 2

6.5.2.1 F0 slope

To better understand how pitch slope varies as a function of Language background, Sentence Type, and Focus structure, we constructed a GLMM targeting the F0_slope of T2 in the second syllable position. Model structure included a full factorial three-way interaction among fixed effects, alongside random intercepts for Speaker and OtherTone. The significance of both random intercepts was confirmed, justifying their inclusion in the model.

Analysis of variance with Satterthwaite's approximation revealed that all three main effects (Language, Sentence Type and Focus), as well as the three-way interaction were significant. These effects were retained in the final model after backward selection, with no simplification of the fixed-effects structure supported by the data.

The fixed-effect estimates from the final model provide key insights into how tonal slope is modulated. The intercept for CH speakers in question.on-focus conditions was positive (β=0.148, p<.001), consistent with the rising nature of T2. A flattening of slope was observed in statement and post-focus contexts, both associated with significant decreases in slope (β=−0.13, p<.001). The Language.Focus interaction was significant and positive (β=0.15, p<.001), indicating that IT did not reduce slope in post-focus conditions as much as CH. Similarly, the Language.Sentence Type interaction (β=0.09, p=.01) suggests that IT differentiate less clearly between statements and questions compared to CH. The three-way interaction was also significant (β=−0.11, p=.039), reflecting that CH speakers reduce their slope the most in statement post-focus contexts, a pattern not fully replicated by learners. This three-way interaction supports the hypothesis that CH integrate tonal and intonational cues in a synergistic manner, whereas IT tend to preserve tone shape across contexts, showing limited pragmatic modulation.

EMMs were extracted for each Language.Sentence Type.Focus cell, and visualized with confidence intervals using a bar plot. The graphical summary reveals a clear asymmetry between native and learner groups: CH exhibit the steepest T2 slopes in question-on-focus contexts, followed by progressive flattening in statement and post-focus positions; in contrast, IT learners' slopes remain relatively stable across conditions, with smaller differences between pragmatic contexts. This pattern suggests that while learners may approximate the pitch shape of T2, they do not implement the PFC or statement-induced flattening observed in native speakers.

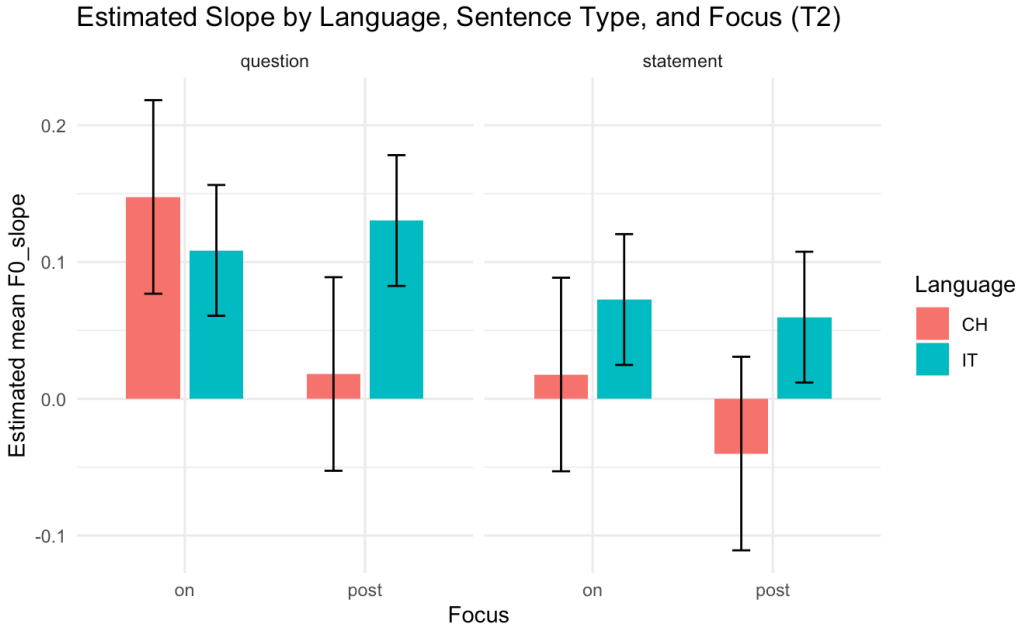


Figure 114 Estimated Slope by Language, Sentence Type, and Focus (T2)

This analysis of F0_slope reinforces the findings from the smooth contour modeling, providing converging evidence that native speakers exhibit prosodically conditioned tonal adjustments. Italian learners, while capable of producing rising contours characteristic of T2, fail to fully implement these prosodic adjustments. The statistically significant three-way interaction among Language, Sentence Type, and Focus confirms that such modulation is context-sensitive.

6.5.2.2 F0 max

In addition to examining tonal slope, the current analysis sought to explore how the maximum pitch height (F0_max) of Mandarin T2 varies as a function of Language background, Sentence Type, and Focus structure. Since T2 is canonically realized as a rising contour, its

peak pitch offers a crucial prosodic marker potentially affected by both linguistic status (L1 vs. L2) and pragmatic context.

To this end, a GLMM was fitted, predicting F0_max as a function of the three-way interaction among Language, S.Type, and Focus, with random intercepts for Speaker and OtherTone.

Model comparison confirmed the necessity of including both random intercepts: removing either Speaker or OtherTone significantly worsened model fit (Speaker: $\chi 2=24.95$, $p<.0001$; OtherTone: $\chi 2=5.90$, $p=.015$), reinforcing the view that inter-speaker variability and tonal context meaningfully shape pitch outcomes.

Analysis of variance with Satterthwaite's approximation on the full model revealed significant main effects for S.Type ($p < .001$) and for the Lang.S.Type interaction ($p<.001$), while Focus yielded a trend-level effect ($p = .17$). Crucially, the three-way interaction did not reach significance ($p = .88$), suggesting that maximum pitch height is not jointly modulated by all three factors. Stepwise model selection confirmed that the best-fitting model retained Lang.S.Type and S.Type.Focus interactions, while removing the three-way interaction and the Lang.Focus term.

The summary of the final model showed several key findings. The intercept estimate corresponds to CH speakers' F0_max in the baseline condition (question, on-focus), establishing a relatively high peak. Compared to the intercept, IT exhibited significantly lower maximum pitch across conditions ($p = .030$), suggesting a generally compressed pitch range or more conservative pitch targets in their T2 productions.

Statements were associated with significantly lower F0_max than questions ($p < .001$), further proving intonational flattening in declarative contexts. Post-focus positions were also characterized by reduced pitch peaks ($p = .049$), consistent with expectations of PFC. Crucially, the Lang.S.Type interaction was highly significant ($p < .001$), revealing that the difference between questions and statements was attenuated in IT relative to CH.

The SentenceType.Focus interaction was also significant ($p = .016$), suggesting that the effect of PFC was stronger in statements than in questions.

EMMs were extracted for each Language.SentenceType.Focus cell and visualized using grouped bar plots with 95% confidence intervals. The graphical output reveals a clear pattern: CH demonstrate marked F0_max differences between questions and statements, especially in on-focus positions, where pitch reaches its highest point. IT, by contrast, show a flatter profile, with notably lower F0_max values and a reduced distinction between sentence types. In post-

focus contexts, both groups lower their pitch peaks, but the reduction appears more systematic in CH, suggesting that PFC is not consistently implemented by IT (see Fig. 115).

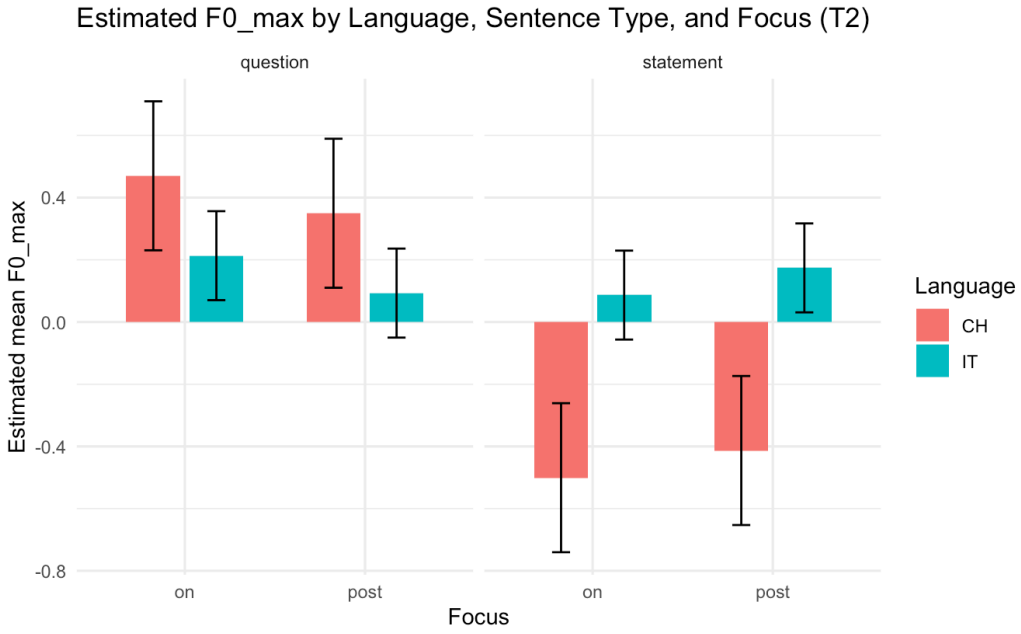


Figure 115 Estimated F0_max by Language, Sentence Type, and Focus (T2)

Taken together, the analysis of F0_max reveals that peak pitch height in T2 production is systematically influenced by sentence type and focus structure, and that these effects interact differently across native speakers and learners. While learners show some sensitivity to post-focus and sentence type variation, their overall pitch range is narrower and less dynamic. The significant Lang.S.Type interaction suggests that learners may preserve the tonal category of T2 but struggle to superimpose intonational contours, particularly in declarative sentences.

6.5.2.3 F0 min

We further examined minimum pitch (F0_min) – the lowest point of the tonal contour, potentially indexing phonetic realization of the rising onset and the speaker's pitch range floor. A GLMM was fitted with Language, Sentence Type, and Focus as fixed effects and random intercepts for Speaker and OtherTone. Model comparison indicated that Speaker was the only significant random effect ($\chi2=41.95$, $p<.0001$), whereas OtherTone provided only marginal contribution ($p=.0525$). Consequently, the model was refitted with Speaker only as a random intercept to avoid overparameterization. Language proved to be the only significant fixed effect ($F(1,73.73)=29.90$, $p<.0001$): neither Sentence Type, Focus, nor any of their interactions reached significance, suggesting that minimum pitch values were invariant across pragmatic

conditions but systematically varied across speaker populations. The backward model selection confirmed this result, arriving at a final minimal model including F0_min~Lang+(1|Speaker).

The model summary showed a robust effect of Language on F0_min. The intercept (β=−1.20, p<.001) corresponds to CH speakers' minimum pitch level. Compared to the intercept, IT exhibited significantly higher minimum pitch values (β=+0.60, p<.001), indicating a misalignment with native-like low starting points in T2 production.

This upward shift in pitch floor suggests a restricted pitch range, or perhaps more specifically, a deficiency in executing low pitch targets of rising tones – potentially due to both physiological constraints and phonological mapping issues in L2 tonal categories.

EMMs were extracted and plotted, providing evidence for a clear, statistically significant difference between CH and IT speakers. Native speakers' F0_min was markedly lower than that of learners, highlighting a persistent cross-linguistic divergence in the low tonal anchoring of T2.

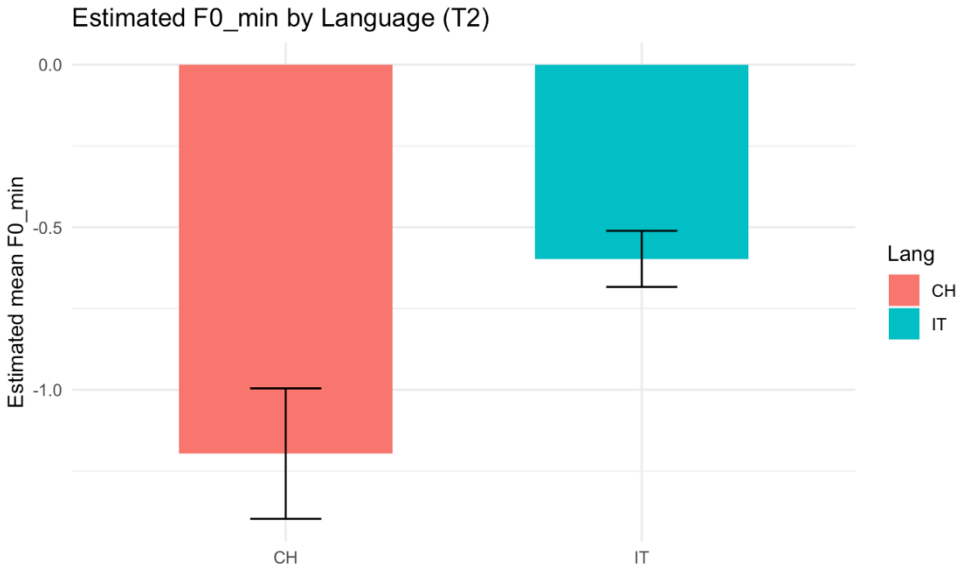


Figure 116 Estimated F0_min by Language (T2)

The analysis of F0_min in T2 production reinforces the notion that low pitch targets pose a specific challenge for L2 learners, who often underutilize the lower end of their pitch range. Unlike pitch peak (F0_max), which may be reached more easily through general intonational strategies or mimicry, reaching a sufficiently low F0 appears to demand more precise motor control and tonal encoding. The absence of modulation by sentence type or focus structure suggests that this is a lexical-level limitation, rather than a discourse-level one.

6.5.2.4 F0 range

To investigate how pitch range – a proxy for tonal dynamism and prosodic expressivity – varies in the production of T2, we modeled the F0_range (difference between maximum and minimum F0) as a function of Language, Sentence Type, and Focus condition, with Speaker and OtherTone as random intercepts. Model comparison using likelihood ratio tests showed that only Speaker significantly contributed to model fit ($\chi 2=26.11$, $p<.0001$), while OtherTone did not ($p = 1.00$). Backward model selection also identified a reduced model as optimal, excluding Focus as main effect. Therefore, the final model retained only Speaker as a random effect, and preserved only the main effects of Language and Sentence Type, along with their interaction

The model output point to several key findings. The intercept ($\beta=1.58$, $p<.001$) reflects the estimated F0_range for CH in questions, serving as the baseline condition. IT exhibited significantly narrower pitch range overall ($\beta=-0.79$, $p<.001$), indicating a general compression of tonal space across conditions. Statements also had reduced pitch range compared to questions ($\beta=-0.82$, $p<.001$), a pattern observed for both groups but with different magnitudes. Crucially, the interaction between Language and Sentence Type was significant ($\beta=+0.71$, $p<.001$), revealing that the drop in pitch range from question to statement was less pronounced among learners than native speakers.

EMMs were plotted to illustrate these interactions in Fig. 117. The plot clearly shows that native speakers produce substantially wider pitch range in questions than in statements, aligning with typical Mandarin intonation patterns. Learners, while also exhibiting reduced range in statements, do so to a lesser degree. More notably, their range in questions is already lower, suggesting an overall compressed tonal realization, particularly in contexts that would otherwise demand greater prosodic modulation.

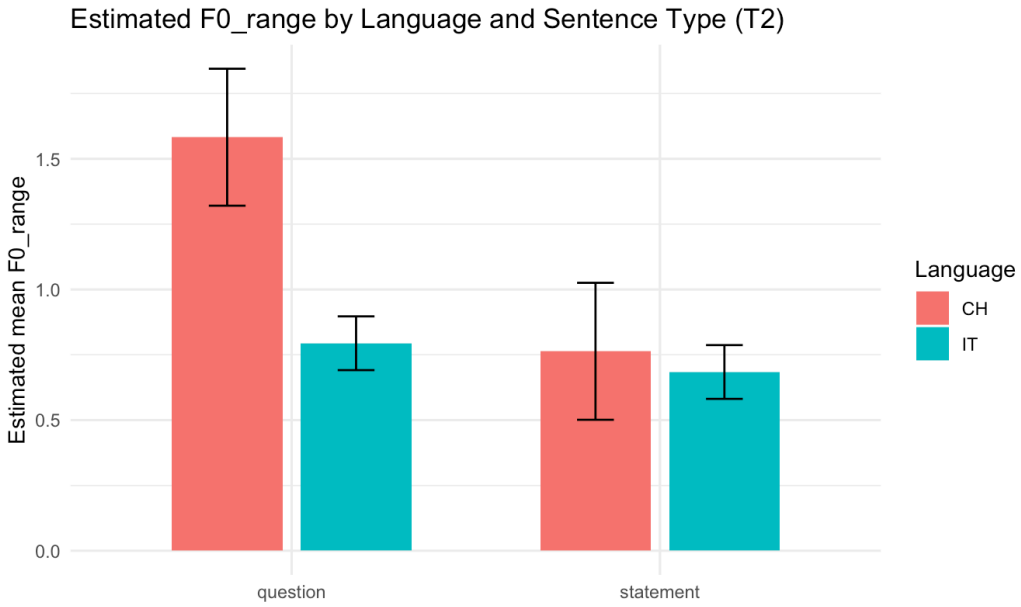


Figure 117 Estimated F0_range by Language and Sentence Type (T2)

These findings offer important insights into L2 tonal phonology and prosody. What is particularly telling is the differential effect of sentence type: native speakers modulate pitch range more dramatically in accordance with sentence type (question vs. statement), a pattern not mirrored by learners. This suggests that intonational cues – layered over lexical tone – may be under-realized or not fully integrated into learners' production. Furthermore, while learners show some sensitivity to sentence-type contrasts, the interaction reveals that their prosodic encoding of discourse type is both attenuated and less dynamic, possibly due to limitations in prosodic planning or pitch control.

The analysis of F0_range in T2 confirms that native-like pitch expansion in tonal realization – particularly for question intonation – is not fully acquired by the L2 group. Although learners do modulate range to some extent, they remain systematically narrower in their pitch excursions and fail to replicate the more expressive pitch scaling observed in native speech.

6.5.3 Italian learner subset

6.5.3.1 Comparing learner-factor models (Proficiency, Musicality, Grade)

To investigate which learner-specific factor most robustly accounts for variation in the F0 trajectories of T2, four mixed GAMM were compared using ML estimation. The baseline model (mSFT2_ml) included only the interaction between Sentence Type and Focus (SFT2),

while three extended models added respectively Proficiency (PSFT2), Musicality (MSFT2), and Grade (GSFT2) as predictor structures.

Pairwise model comparisons revealed distinct patterns. The model incorporating Proficiency (mPSFT2_ml) did not significantly improve over the baseline. The chi-square test for ML scores yielded a non-significant result ($\chi^2(12) = 1.999$, $p = 0.983$), and the AIC difference was minimal ($\Delta AIC = 2.35$), suggesting that proficiency level, at least as operationalized here, does not substantially contribute to variation in T2 realization. In contrast, the model including Musicality (mMSFT2_ml) did show a statistically significant improvement ($\chi^2(12) = 12.62$, $p = 0.014$), with a corresponding AIC reduction of nearly 34 points compared to the baseline. Even more robust was the improvement offered by the Grade-based model (mGSFT2_ml), which demonstrated a highly significant ML score difference ($\chi^2(24) = 30.71$, $p < .001$), and an AIC decrease of 47 points relative to the SFT2-only model.

To confirm these results, a three-way AIC comparison among the best-performing models (mSFT2, mMSFT2, mGSFT2) was conducted and illustrated in Tab. 42 below:

Table 42 Musicality and Grade model comparison with baseline

Model	Degrees of Freedom	AIC
mSFT2_ml	202.42	16511.86
mMSFT2_ml	219.72	16477.93
mGSFT2_ml	220.06	16464.75

The Grade-based model (GSFT2) emerged as the best-fitting model, with the lowest AIC overall, suggesting that it captures the variation in T2 production most effectively.

6.5.3.2 Interaction of Grade, Sentence Type and Focus

To further examine how grade level interacts with sentence structure and information focus in shaping T2 production among Italian learners, we refitted the GAMM incorporating the three-way interaction term; the model included random smooths for Speaker and OtherTone[27].

The parametric component included comparisons between baseline (BA2.question.on) and other interaction levels. Several group-level effects emerged: BA3.question.on showed a significantly higher F0_z intercept relative to the baseline ($\beta = 0.337$, $p = 0.006$), suggesting

[27] The model converged successfully under the fREML criterion. The basis dimension check revealed no undersmoothing problems, with all k-index values around 1.02 and non-significant p-values, indicating that the smooth terms were well-specified for the complexity of the data.

more elevated pitch onset for more advanced learners in interrogative, on-focus contexts; BA2.statement.on and BA3.statement.on also demonstrated significant positive shifts (p = 0.022 and p = 0.046 respectively), whereas MA1.statement.on did not differ from the baseline. Similarly, BA3.statement.post revealed a positive and significant coefficient (β = 0.318, p = 0.010), possibly reflecting higher pitch targets in final positions among the highest-performing subgroup. Other conditions, including MA1 across question and post-focus contexts, did not exhibit significant parametric shifts.

The non-linear time-varying effects captured by smooths over Point revealed additional subtleties in pitch trajectory shaping (see Figg. 118-121). Among question.post contours, BA3 learners showed a clearly significant smooth term (edf = 2.79, F = 5.65, p < .001), indicating robust time-varying modulation of F0 – presumably reflecting a canonical rising pattern of T2. BA2.question.post exhibited a marginal effect (p = 0.08), while MA1.question.post was non-significant.

Smooths for question.on and statement contexts were generally weaker, with most edf values near 1 and non-significant p-values, suggesting less engagement with dynamic tonal targets. Specifically to statement.post, such flattening may index a more native-like tone target undershoot in BA3 and MA1 productions.

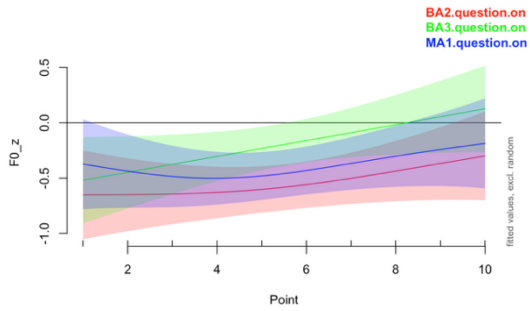


Figure 118 T2 question on-focus production by Grade

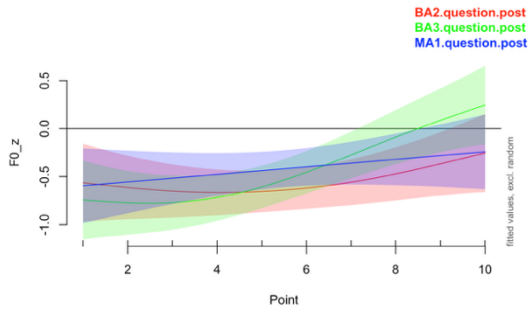


Figure 119 T2 question post-focus production by Grade

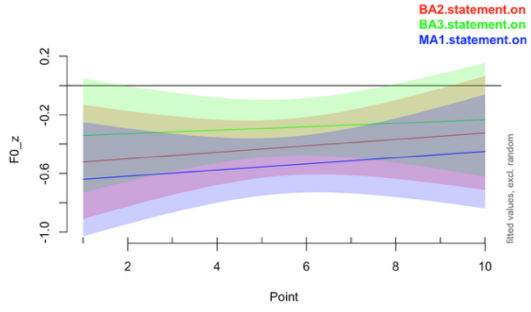


Figure 120 T2 statement on-focus production by Grade

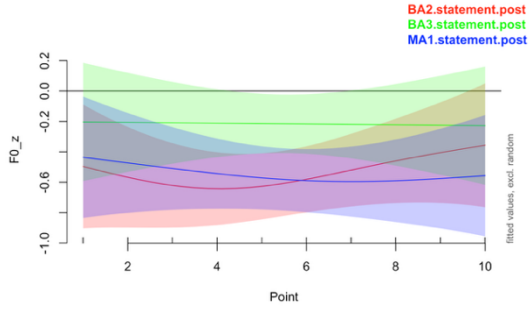


Figure 121 T2 statement post-focus production by Grade

6.5.3.3 Interaction of Musicality, Sentence Type and Focus

To further understand the interindividual variability in T2 realization among Italian learners, we refitted a GAMM that integrated Musicality, Sentence Type, and Focus as interacting predictors.

The model predicted F0_z as a function of the interaction between Musicality (High vs. Low), Sentence Type, and Focus, with time (as Point) modeled via condition-specific smooths; the model also controlled for individual variation and contextual tonal interference using factor smooths for Speaker and OtherTone[28].

The parametric coefficients revealed significant group-level differences between high and low musicality learners. Specifically, learners in the High Musicality group consistently exhibited lower F0 baselines across several conditions. This was particularly marked in question.post (β = -0.178, $p < .001$), statement.post (β = -0.127, $p < .001$), and statement.on (β = -0.123, $p < .001$), suggesting that higher musical aptitude may promote better alignment with native-like PFC and non-interrogative contexts. Conversely, differences between High and Low groups in question.on and statement.on/post conditions were not statistically significant, hinting at more overlap in pitch height or increased variability in these contexts.

The non-parametric smooth terms indicated that in the Low.question.on group, the smooth term was significant (*edf* = 2.93, $p = 0.014$), proving non-linear pitch movement. This may suggest overarticulation or compensatory pitch gestures in the absence of refined prosodic control. For the High.question.post group, the smooth was also significant (*edf* = 2.69, $p = 0.004$), with a clearly modulated pitch contour. Other smooths were non-significant and often approximated linear functions (edf ≈ 1), indicating flatter pitch trajectories or reduced pitch modulation (see Figg. 122-125).

[28] The model converged under the fREML criterion and successfully passed the gam.check() procedure. All k-index values equaled 1.00 with p-values well above the .05 threshold, indicating that the chosen basis dimensions were appropriate and that the model avoided underfitting.

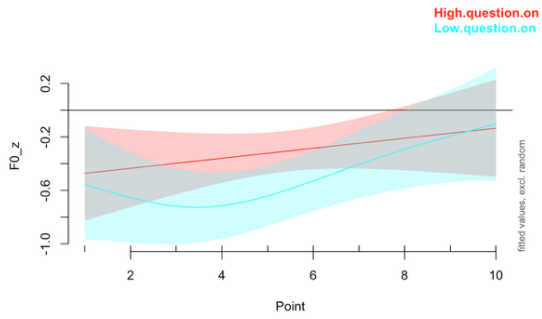


Figure 122 Tone 2 question on-focus production by Musicality

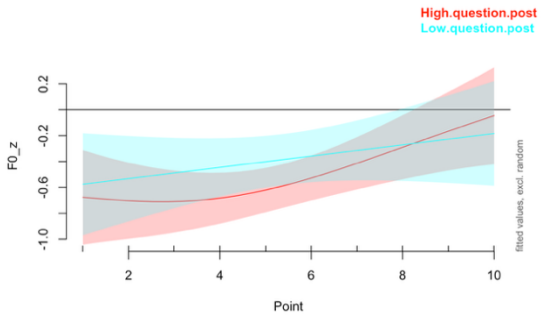


Figure 123 Tone 2 question post-focus production by Musicality

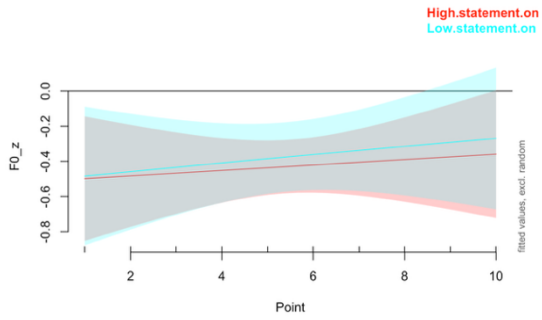


Figure 124 Tone 2 statement on-focus production by Musicality

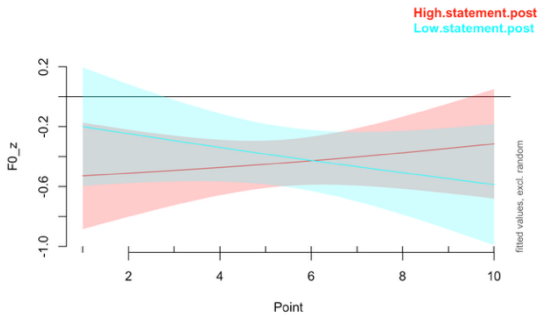


Figure 125 Tone 2 question post-focus production by Musicality

6.5.4 Interim summary on T2

The results from the T2 subset offer compelling insights into the tonal-prosodic integration among Italian learners' L2 Mandarin production. Despite T2's relatively transparent phonological identity – a rising contour – it emerges as a particularly revealing diagnostic of learners' capacity to interface lexical tone with sentence-level prosody.

Native speakers modulate T2 pitch height and contour in a context-sensitive manner, implementing PFC especially in statements and post-focus environments. This is evident both in flattened GAMM smooths and significantly reduced F0 parameters (e.g., F0_z, F0_slope, and F0_max). Crucially, in post-focus contexts, native speakers often undershoot the canonical T2 rise, producing either compressed or flat contours. In contrast, Italian learners show a markedly reduced pitch modulation capacity: learners' productions remain closer to the citation form across conditions, maintaining rising contours and elevated pitch levels even in contexts that natively favor tonal reduction. These patterns are most salient in the F0_slope and F0_max analyses, where learners fail to replicate the tonal flattening typical of native speakers in declarative and post-focus environments. Moreover, learners consistently produced higher F0_min values, indicating a restricted pitch floor, which limits full realization of rising contours

and likely reflects both physiological limitations and phonological under-specification of tonal anchoring.

Notably, intonational distinctions between sentence types were also attenuated in the Italian learners' group. While native speakers expanded pitch range and slope in interrogatives, learners exhibited compressed F0_range and reduced sentence-type contrasts, suggesting underdeveloped intonational layering over lexical tone.

Individual differences within the IT group shed further light on L2 tone acquisition trajectories. General language proficiency, though predictive for T1 modulation, had minimal explanatory power for T2. Instead, academic grade level and musical aptitude emerged as stronger predictors of native-like pitch realization in these experimental conditions. Learners at higher academic levels (especially BA3) showed more nuanced pitch shaping, particularly in rising contours of question-final positions. Likewise, musically trained learners exhibited greater pitch control, especially in non-interrogative contexts, and implemented more native-like compression post-focus.

Taken together, these findings underscore the complex interaction between tone, intonation, and focus structure in L2 Mandarin. While learners may acquire the phonemic category of T2, they often fall short in implementing its pragmatic flexibility. The overapplication of citation-form contours across sentence types – especially the tendency to preserve rising slopes where native speakers compress or flatten – suggests that tone categories are acquired before their prosodic modulation is mastered.

The present study confirms that prosodic integration – especially under information structure and sentence type pressures – is a persistent bottleneck in L2 tone acquisition. Across all four parameters, native speakers show systematic modulation of T2 properties based on sentence type and focus: rising slopes and wide pitch ranges in questions, flatter slopes and compressed ranges in statements, and post-focal reductions where expected. By contrast, Italian learners exhibit more neutralized or underdifferentiated contours, failing to adjust tonal properties to suit higher-level prosodic contexts.

This suggests that although learners may attain phonological accuracy at the segmental or lexical level, they often lack prosodic sensitivity and flexibility, especially for tones embedded in complex intonational contexts.

6.6 Tone 4 Subset Analysis

6.6.1 Interaction of Language, Sentence Type, and Focus

To investigate the production of T4 in disyllabic targets, we fitted a GAMM that incorporated a three-way interaction between Language, Sentence Type, and Focus.

To operationalize the model, a new interaction variable was created by combining the three factors, resulting in eight unique conditions (e.g., CH.question.on, IT.statement.post, etc.). Each of these was associated with its own smooth term, capturing dynamic F0 movements over time; the model also included random factor smooths for Speaker and OtherTone[29].

The smooth terms provided critical insights into the temporal dynamics of T4 production. Native speakers displayed complex F0 trajectories in most conditions, especially in CH.question.on (edf = 4.14, p < .001) and CH.question.post (*edf* = 3.04, *p* < .001), reflecting the characteristic convex shape of T4, which typically starts high and falls sharply.

Italian learners showed more modest contour modulation, with flatter curves in IT.question.on (edf = 3.19, p = .030) and IT.statement.post (edf = 3.87, p < .001), but also displayed significant shaping in IT.statement.on (edf = 4.47, p < .001). These results suggest that non-native speakers modulate pitch directionally on statement conditions, but may lack the degree of pitch excursion or sharpness in falling movement typical of native-like T4. What is particularly telling here are the question conditions, which – despite exhibiting some degree of tonal modulation – remain considerably flatter than the statement conditions, regardless of focus. This general lowering and flattening of the tone contour seems to be caused by the question condition itself. Indeed, although learners partially mastered the tone in isolation (as demonstrated in Study 1, § 4), they still tend to produce a falling contour in statements and focus contexts. Such a pattern may reflect a degree of negative prosodic transfer from Italian, through the approximation of a L* LH% boundary tone used to mark interrogativity (see Figg. 126-129).

In CH.statement.post, the smooth was almost linear (*edf* = 1.00), though still significant (*p* = .002). The Speaker and OtherTone smooths were also highly significant (p < .001), reaffirming the influence of individual pitch settings and tonal coarticulation on tonal realization.

---

[29] The model converged successfully using the fREML criterion and passed all diagnostic checks (k-index = 1.02 across smooths, all p > 0.90), suggesting an appropriate basis dimension and stable estimation. The model explained 40% of the total deviance, with an adjusted R^2 of 0.381, indicating a robust fit given the complexity of the tonal-prosodic interface and inter-speaker variability.

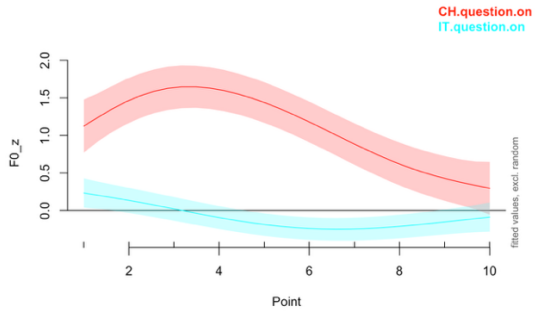


Figure 126 Tone 4 question on-focus production by Language

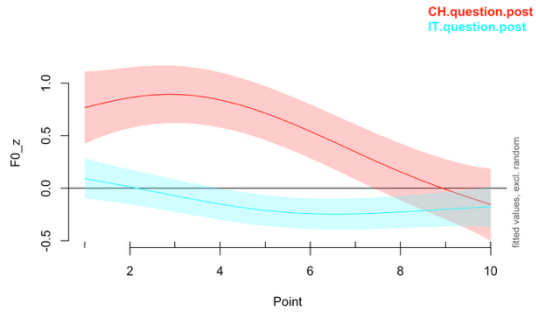


Figure 127 Tone 4 question post-focus production by Language

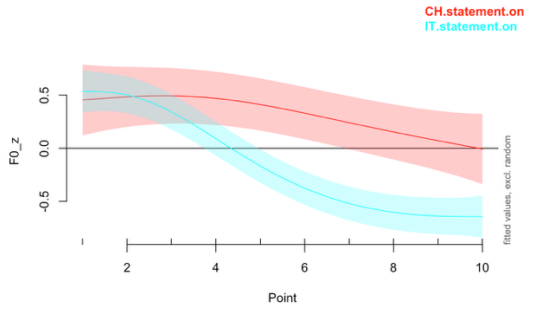


Figure 128 Tone 4 statement on-focus production by Language

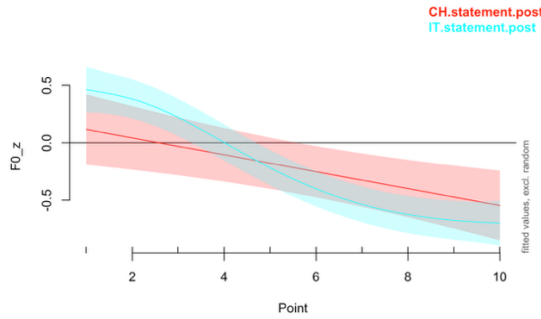


Figure 129 Tone 4 statement post-focus production by Language

To directly assess cross-linguistic differences within each condition, we refitted the model using three separate fixed factors (Lang, S.Type, Focus) and extracted EMMs using the *emmeans* package.

The pairwise comparisons revealed that CH consistently produced higher F0_z values across all conditions, with three of the four contrasts reaching statistical significance, as shown in Tab. 43:

Table 43 Pairwise comparison between Language groups

Sentence Type	Focus	Contrast (CH - IT)	p-value
Question	On	1.536	< .0001
Statement	On	0.653	< .0001
Question	Post	0.863	< .0001
Statement	Post	0.100	0.4143

The absence of a significant group effect in the statement-post condition indicates that IT approximate native-like pitch levels more closely when producing falling tones in post-focus contexts, where T4 is typically undershot. Nevertheless, examination of the predicted F0 contours reveals that IT still preserve a degree of high-convex trajectory, characteristic of the citation form, rather than adopting the compressed or flattened realization commonly observed in CH productions. This discrepancy suggests a continued reliance on segmental tone targets in isolation, with limited evidence of the prosodic neutralization processes expected in post-focal environments.

The EMM bar plots visualized these differences across conditions, clearly showing the systematic pitch height reduction in Italian learners' T4 production (see Fig. 130). Notably, the greatest discrepancy occurred in question-on contexts, likely due to the compounding effects of focus prominence and interrogative raising in pitch height in Chinese – prosodic phenomena that may trigger negative transfer in L2 learners productions.

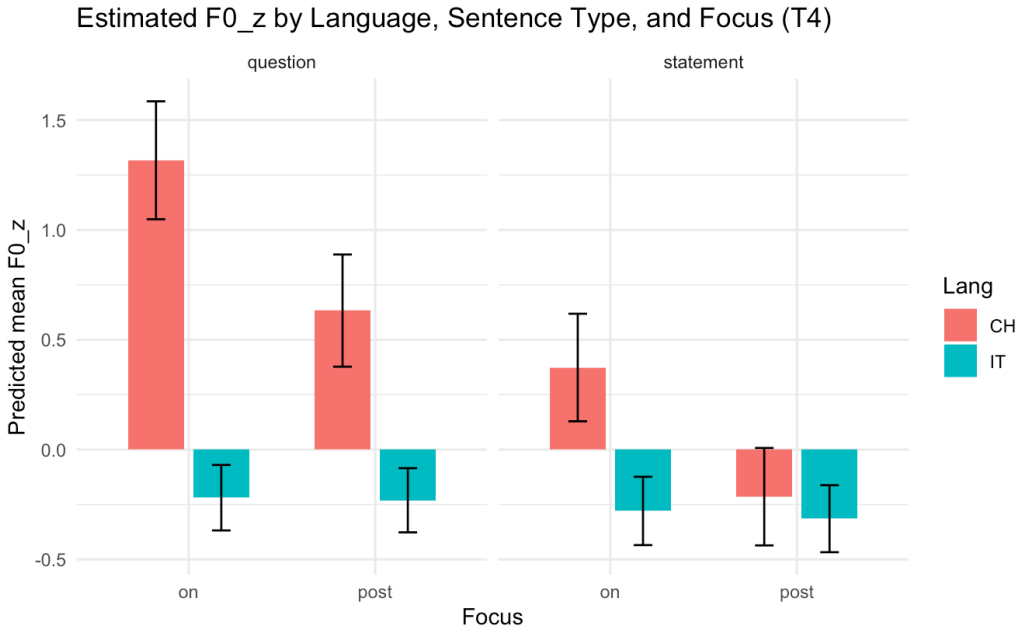


Figure 130 Estimated F0_z by Language, Sentence Type, and Focus (T4)

This model of T4 production confirms that non-native learners significantly underperform in replicating native-like pitch contours, especially in interrogative positions. While learners exhibit some degree of contextual pitch modulation, their realizations generally feature flattened contours and compressed pitch ranges, leading to possible perceptual ambiguity or loss of prosodic nuance.

6.6.2 Analysis on Curve Parameters for Tone 4

6.6.2.1 F0 slope

To complement the contour-based analysis of T4, we conducted a GLMM analysis on the F0_slope of the final syllable. The primary goal was to examine whether pitch falling behavior – crucial to the realization of T4 – varies as a function of Language, Sentence Type, and Focus.

We initially fitted a full model including random intercepts for both Speaker and OtherTone. A likelihood-ratio test revealed that Speaker accounted for a significant portion of the variance ($p < 2e\text{-}16$), whereas OtherTone did not ($p = 0.21$). The model was therefore refitted with Speaker as the only random effect, improving both interpretability and parsimony.

An initial Type III ANOVA on the fixed effects with Satterthwaite's approximation identified a significant main effect of Sentence Type ($p = .021$), as well as a highly significant Lang.S.Type interaction ($p < 2e\text{-}11$). No significant effects of Focus or any higher-order interactions were found. This led to a stepwise model reduction, which converged on a final model retaining only Language, Sentence Type, and their interaction, with Speaker included as a random effect.

The intercept reflects the F0_slope in the CH.question condition, which is negative as expected for a falling tone. The positive main effect of Sentence Type indicates shallower falls in statements compared to questions, but this pattern reversed for IT, as indicated by the strong negative interaction. In other words, while CH maintain steeper F0 falls in questions, IT learners show flatter slopes in questions and steeper slopes in statements, reflecting a potential misalignment with native prosodic norms.

EMMs were extracted for each Language.Sentence Type condition and visualized via grouped bar plots in Fig. 131 below.

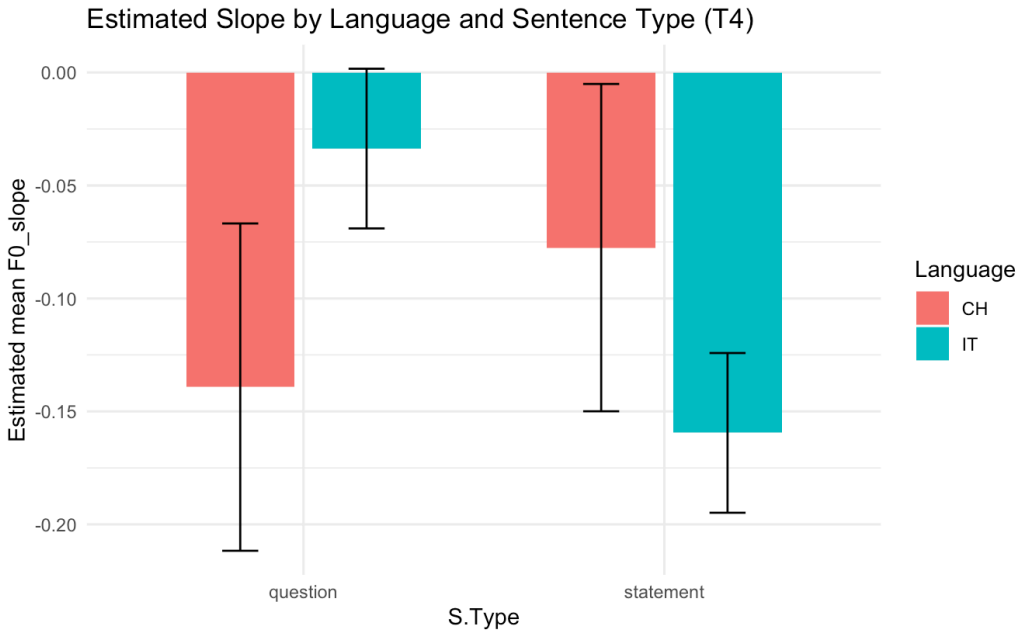


Figure 131 Estimated Slope by Language and Sentence Type (T4)

In the native group, questions exhibited steeper negative slopes, consistent with a canonical T4 falling contour enhanced by interrogative intonation. In contrast, Italian learners exhibited an opposite trend, with less steep F0 slopes in questions and relatively sharper falls in statements, suggesting either a transfer of L1 intonational patterns or a misinterpretation of pitch prominence in question contexts.

This asymmetry was statistically robust and mirrors the findings from the GAMM smooths of the full contour, which showed that IT tend to flatten or reduce pitch excursion in contexts requiring prosodic emphasis (e.g., questions).

The slope analysis corroborates the contour-based findings: T4 production is significantly modulated by both sentence type and language background, with non-native learners failing to align their prosodic patterns with native expectations. Specifically, while CH exploit slope dynamics to reinforce pitch cues in questions, IT appear to invert or neutralize this cue, resulting in prosodic mismatches that may affect falling tone intelligibility.

6.6.2.2 F0 max

To further examine how prosodic realization of T4 varies across speakers and contexts, we analyzed F0_max on the final syllable using a GLMM. The analysis aim to observe how Language, Sentence Type, and Focus modulate peak pitch height.

A likelihood-ratio test suggested that both Speaker ($p < 2.2e-16$) and OtherTone ($p = 7.28e-06$) contributed significantly to the model fit. These random effects were retained in all subsequent modeling.

A Type III ANOVA on fixed effects with Satterthwaite approximation justified retaining the two-way interactions Lang.S.Type and Lang.Focus, while higher-order terms were dropped during stepwise model reduction.

The significant main effect of Language indicates that, overall, IT produced lower F0 maxima compared to CH. Similarly, statements and post-focus conditions were associated with lower peak pitch values, consistent with known focus-marking patterns in Mandarin Chinese (see § 2.6.2).

Crucially, the Lang.S.Type interaction shows that while CH consistently produced higher pitch peaks in questions than in statements, this contrast was reduced or reversed in IT. The Lang.Focus interaction reveals that CH also modulated peak F0 more sharply in response to focus structure, a sensitivity not equally matched by the IT group.

EMMs from the final model were plotted across Sentence Type and Focus conditions. The plots revealed that native speakers modulated peak F0 reliably across both sentence type and focus, indicating higher F0_max in questions and on-focus positions; Italian learners, in contrast, produced attenuated pitch peaks overall, and their intonational modulations were less differentiated, particularly in post-focus contexts. These findings align closely with the slope and contour analyses, where IT often failed to fully realize pitch dynamics associated with tonal and intonational contrasts.

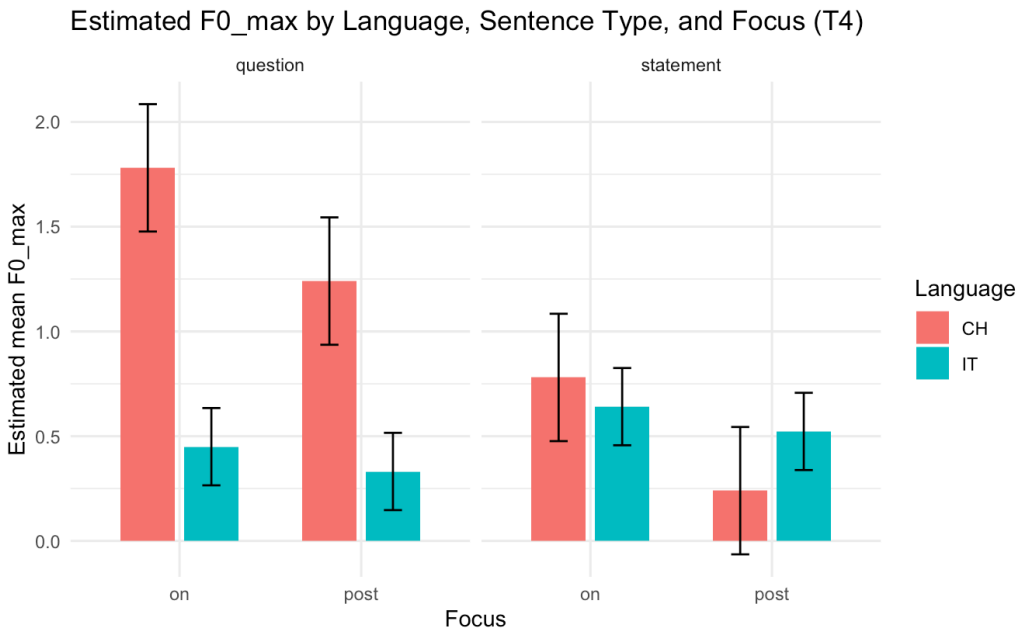


Figure 132 Estimated F0_max by Language, Sentence Type, and Focus (T4)

The analysis of F0_max reveals strong effects of language background, sentence type, and focus, with clear differences in how native and non-native speakers employ pitch maximum height. Italian learners exhibit reduced dynamic range and weaker sensitivity to discourse-level prosodic cues, particularly in sentence type and PFC. These results suggest that L2 tonal acquisition challenges extend beyond contour shape to include fine-grained pitch control, especially in pragmatically rich prosodic environments.

6.6.2.3 F0 min

To further explore prosodic realization in the production of T4, we analyzed the minimum fundamental frequency (F0_min) of the syllable. This analysis complements the investigation of F0_max and F0_slope, enabling a full picture of pitch excursion.

A likelihood ratio test indicated that Speaker significantly contributed to the model ($p < .00001$), while OtherTone did not ($p = .159$). Therefore, Speaker was retained as the sole random effect in the final model.

Significant interactions and main effects were revealed for Focus, Lang.S.Type, and Lang.Focus; S.Type (main effect) was marginal ($p = .071$), and no higher-order interaction reached significance. Stepwise model reduction yielded the final model, which retained both two-way interactions but dropped the three-way interaction and the S.Type.Focus term.

Results show that IT displayed lower F0_min values than CH across conditions ($\beta = -0.472$), suggesting reduced pitch range or floor in tone realization.

Across both groups, post-focus conditions were associated with significantly lower F0_min values. This reflects a canonical PFC, more prominent in CH, but at least partially present in IT productions. Statements exhibited lower minima than questions (β = -0.363), in line with the declarative fall typical of T4, especially when not overridden by interrogative rise or prosodic focus.

The Lang.S.Type interaction supported the view that Italian learners' pitch floor in statements is less reduced compared to native speakers, suggesting flattening of the tone shape in statements.

The Lang.Focus interaction indicates reduced PFC among Italian speakers, supporting the hypothesis that focus-sensitive prosody is not yet fully acquired. These patterns suggest that non-native speakers do not yet fully master the low pitch targets expected in T4 realizations, especially in post-focus and statement contexts, where native speakers show pronounced lowering of F0_min.

EMMs were visualized using bar plots across Focus and Sentence Type conditions for both language groups in Fig. 133. CH produced markedly lower F0_min in post-focus positions, especially in statements, consistent with native-like focus marking and falling tonal contours.

IT, although still exhibiting some focus-related pitch lowering, demonstrated less compression overall, particularly in the statement-post-focus condition.

These results support related findings from the F0_max and F0_slope analyses in this work, reinforcing the view that pitch excursion control is a key challenge in L2 prosody acquisition.

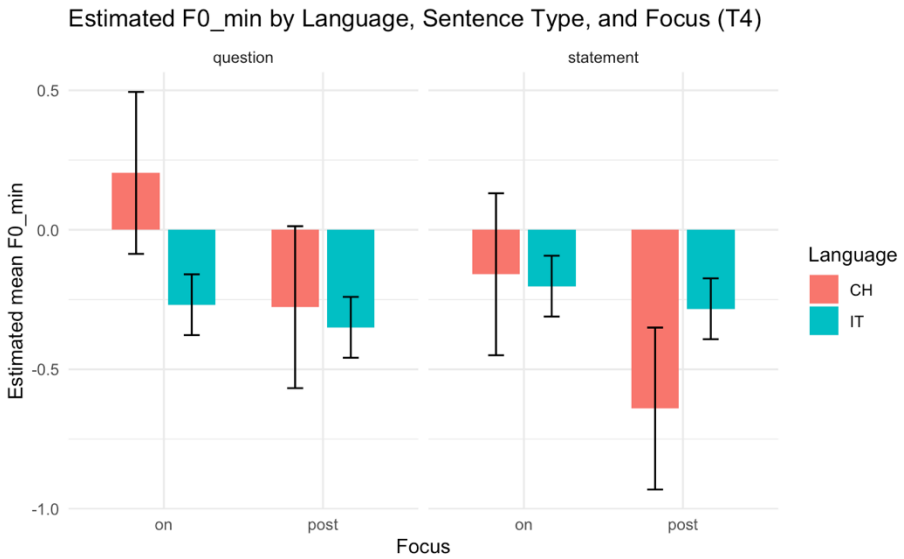


Figure 133 Estimated F0_min by Language, Sentence Type, and Focus (T4)

This analysis of F0_min provides further evidence of cross-linguistic divergence in prosodic realization. While Italian learners demonstrate a degree of sensitivity to focus structure and sentence type, they fall short of native-like pitch floor modulation, especially in low-tone targets and prosodic compression contexts. These results, when viewed alongside max pitch and contour dynamics, emphasize a systematic limitation in pitch range control among learners.

6.6.2.4 F0 range

To investigate the extent to which speakers modulate pitch excursion when producing T4 syllables, we modeled the F0_range – operationalized as the numerical difference between the syllable's maximum and minimum F0.

We first assessed the contribution of random effects in a GLMM that included Speaker and OtherTone. The model comparison using likelihood ratio tests showed a significant contribution for Speaker ($p < .00001$), but not for OtherTone ($p = 1$). Thus, only Speaker was retained in the final model. Initial Type III ANOVA identified Language, Sentence Type, and their interaction as significant predictors of F0_range. All other main effects and interactions, including Focus, were non-significant ($p > .24$). Accordingly, stepwise model selection yielded a parsimonious model that included Language, Sentence Type, and their interaction as fixed effects, with Speaker entered as a random effect.

Overall, IT produced smaller pitch excursions than CH. The negative coefficient ($\beta = -0.848$) reflects a compressed pitch range, indicative of undershoot in tonal realization or reduced motoric control over pitch modulation. Across both groups, statements were associated with a narrower F0_range compared to questions ($\beta = -0.637$). This is consistent with the declarative intonation pattern in Mandarin, where falling contours may be realized with less dynamic pitch movement in statements.

Interestingly, the significant positive interaction ($\beta = +0.762$) suggests that the difference in F0_range between questions and statements is smaller in IT productions. That is, while CH dynamically adjust pitch range depending on sentence type, IT exhibit a reduced tonal flexibility across discourse contexts.

The results offer strong evidence that learners show restricted control over pitch excursion, with this limitation most pronounced in syllables involving T4. Furthermore, their reduced sensitivity to sentence type in terms of pitch range manipulation suggests a lack of integration between lexical tone and intonational structure, a pattern also observed in the analyses of F0_max, F0_slope, and F0_min.

The EMMs plot clearly illustrates the following effects: CH exhibit greater pitch range, especially in question contexts, consistent with high tonal prominence and expressive intonation; whereas, IT produce flatter tone contours, with comparatively smaller differences in pitch range between questions and statements, reinforcing the interpretation of tonal compression and prosodic rigidity (see Fig. 134).

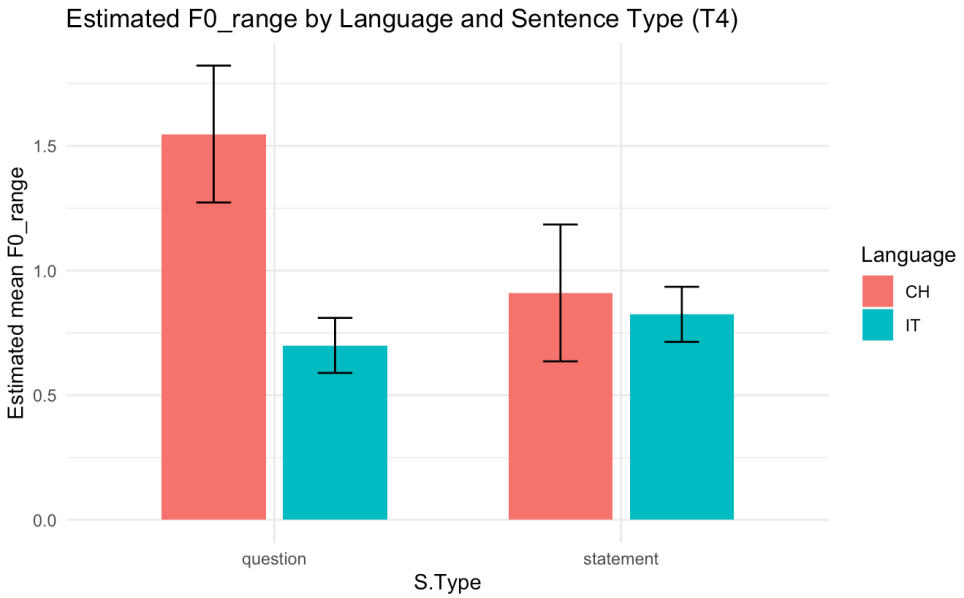


Figure 134 Estimated F0_range by Language and Sentence Type (T4)

The F0_range analysis enriches our understanding of L2 tonal production by capturing tonal dynamism. The results suggest a systematic limitation in Italian learners' control of pitch range, which becomes especially evident in their adaptation to sentence-level prosodic structures. This adds to the accumulating evidence that L2 tone acquisition is not only a matter of hitting target tones but also involves mastering the pragmatic flexibility and prosodic plasticity characteristic of native-like tonal prosody.

6.6.3 Italian learner subset

6.6.3.1 Comparing learner-factor models (Proficiency, Musicality, Grade)

To assess the contribution of individual-level factors to the realization of Mandarin T4 in sentence-final position among Italian learners, we constructed a series of GAMMs and compared their fits under Maximum Likelihood estimation. The baseline model (mSFT4_ml) included only the interaction between sentence type and focus (SFT4) as a fixed predictor, along with smooths over pitch point by condition, speaker, and co-occurring tone. We then

compared this reference model to extended models that incorporated one of three speaker-specific variables: Proficiency (mPSFT4_ml), Musicality (mMSFT4_ml), and Grade (mGSFT4_ml). These comparisons allowed us to isolate the predictive utility of each variable in accounting for variation in normalized pitch (F0_z).

The comparison between the baseline model and the proficiency-augmented model yielded a statistically significant improvement in model fit. Specifically, the likelihood ratio test indicated a χ^2 difference of 15.54 with 12 degrees of freedom ($p = 0.002$), and the AIC decreased by approximately 9.51 points in the extended model. This suggests that speaker Proficiency, as operationalized here, provides a modest but meaningful improvement in capturing T4 pitch trajectories.

By contrast, the inclusion of Musicality in the model did not yield any statistically significant improvement over the baseline. The extended model including Musicality (mMSFT4_ml) exhibited a non-significant χ^2 difference ($p = 0.497$), and the AIC difference favored the simpler SFT4 model. This lack of effect stands in contrast to earlier results observed for T2, where Musicality had emerged as a significant modulator of pitch production. The discrepancy here may reflect the different pitch dynamics of T2 and T4: while T2 requires gradual pitch rise, which may benefit from musical pitch-tracking skills, T4 is characterized by abrupt pitch drop, possibly less dependent on fine auditory resolution and more susceptible to global prosodic factors like phrase-final declination. These results suggest that Musicality, as operationalized here, is not a generalizable predictor of tone production proficiency, and its effects may be limited to tonal categories with more complex or less canonical contours.

The most striking result emerged from the model incorporating Grade level (mGSFT4_ml). This model outperformed both the baseline and all other extended models by a wide margin. Compared to the baseline, it yielded a χ^2 difference of 97.66 with 24 degrees of freedom ($p < 2e\text{-}16$), and a dramatic AIC reduction of 156.05 points, underscoring the robustness of the improvement. In fact, when directly compared via AIC, it emerged as the best-fitting model, as reported in Tab. 44 below.

Table 44 Grade and Proficiency model comparison with baseline

Model	AIC
mGSFT4_ml	16510.15
mPSFT4_ml	16656.69
mSFT4_ml	16666.20

6.6.3.2 Interaction of Grade, Sentence Type and Focus

To further investigate the effect of academic progression on the production of Mandarin T4 by Italian learners, we refitted a GAMM incorporating grade level, sentence type, and focus condition in a three-way interaction. This was operationalized via the GSFT4 variable, which encoded combinations of Grade, Sentence type, and Focus. The dependent variable was F0_z over time (Point), with random smooths for Speaker and OtherTone to account for repeated measures and co-articulatory tonal context.

The model was refitted using the fREML method in the *bam()* function from the *mgcv* package, with a tensor-product basis for by-condition smooths. The smooths over time were defined per level of GSFT4, allowing for flexible modeling of pitch contours within each condition[30].

Among the parametric terms, only a subset reached statistical significance. Notably, the BA2.statement.on and BA2.statement.post groups showed significantly lower intercept values compared to the baseline (BA2.question.on), with estimates of −0.47 ($p < .001$) and −0.29 ($p<.001$), respectively. These lower intercepts suggest a systematic pitch lowering in BA2 productions in sentence-final statements, consistent with increased pitch compression and/or difficulty in maintaining pitch targets under prosodic pressure. All other parametric contrasts failed to reach significance, indicating that the main variation in tonal shape is better captured by the smooth components rather than levelwise differences in intercepts.

The model revealed significant and distinct pitch contour shapes across grade levels and prosodic contexts, captured by the by-condition smooth terms. Several key patterns emerged. In on-focus questions, smooth terms were significant across all three grade levels. BA2 and BA3 learners exhibited moderately rising F0 contours, whereas MA1 learners produced a smoother and more consistent pitch fall ($F = 12.84$, $p < .001$), approximating native-like T4 trajectories (see Fig. 135). This suggests that more advanced learners possess improved control over pitch realization, even in prosodically marked sentence-final contexts. However, as previously noted in the Grade.Sentence Type.Focus analysis of the T1 subset, such segmental accuracy does not necessarily indicate prosodic flexibility. Notably, MA1 learners' question contours were characterized by a lower pitch onset compared to statements – contrary to the

[30] Model diagnostics indicated a satisfactory fit, with no signs of overfitting or underfitting. The effective degrees of freedom (edf) for the by-condition smooths remained well below the basis dimension ($k = 10$), and gam.check() returned k-index values around 1 for all terms ($p > 0.49$), suggesting no need for increasing complexity. The model explained 35.9% of the deviance, with an adjusted R^2 of 0.338, a substantial improvement over simpler models previously tested (e.g., PSFT4 and MSFT4), and confirms the robustness of Grade level as a predictive variable in shaping T4 production.

native speaker pattern – suggesting a misalignment in question prosody despite accurate tone shape. This observation aligns with the EMMs comparison, which showed persistent differences in F0 height between learner and native productions.

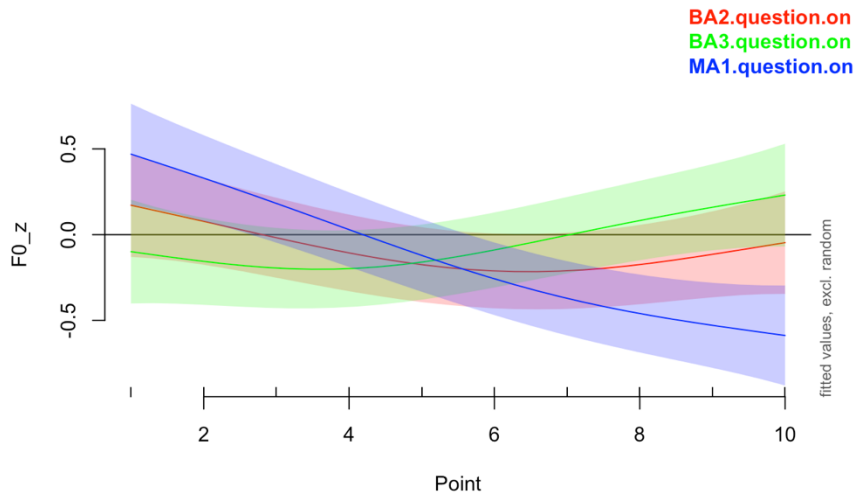


Figure 135 T4 question on-focus production by Grade

In on-focus statements, all groups again displayed significant smooth terms: MA1 had the most constrained smooth (edf = 1.00); in contrast, BA3 had more complex shapes (edf = 2.81), suggesting greater intra-learner variability or difficulty adapting to statement-final prosody (see Fig. 136).

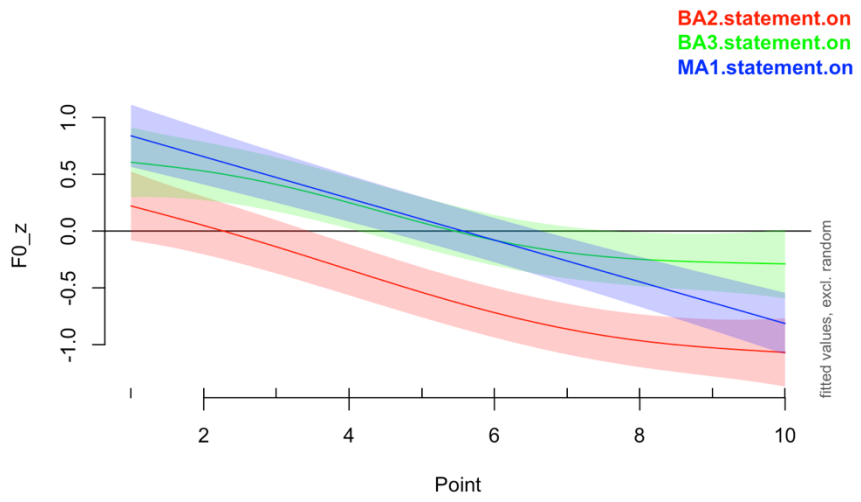


Figure 136 T4 statement on-focus production by Grade

In post-focus questions, only MA1 learners produced a clearly significant smooth ($p < .001$), with a simplified edf of 1.00, suggesting a high level of pitch control and possibly focus-induced deaccentuation (Fig. 137).

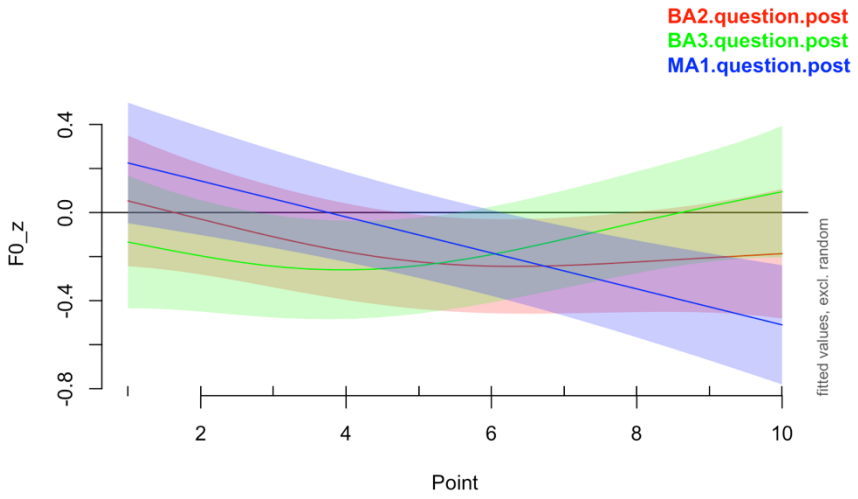


Figure 137 T4 question post-focus production by Grade

In post-focus statements, all three groups showed significant pitch movement: BA3 displayed the most complex contour (edf = 3.01), indicating that even at intermediate levels, learners actively adjust their pitch trajectory in response to post-focal deaccenting, albeit with variability (see Fig. 138).

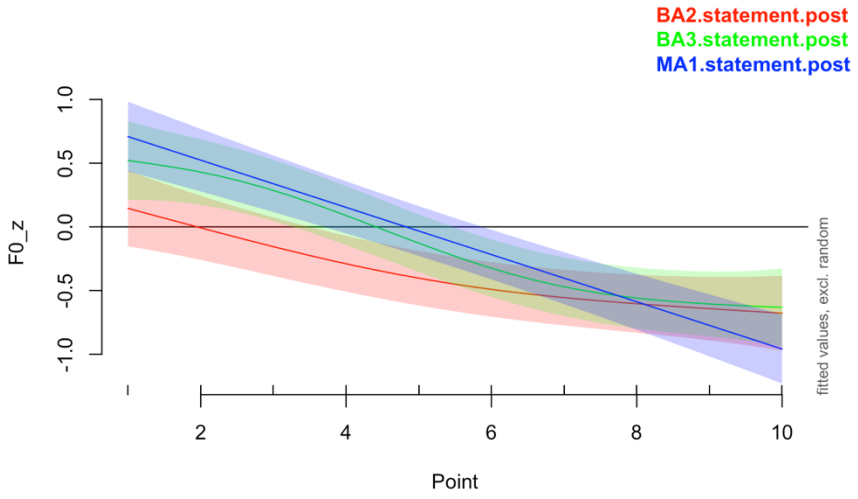


Figure 138 T4 statement post-focus production by Grade

These smooths illustrate that tonal implementation varies not only across sentence conditions but also across academic experience. Notably, MA1 learners generally produce more canonical falling contours, even in prosodically demanding positions (e.g., post-focus), while BA2 learners show compressed or flattened F0 trajectories, consistent with negative L1 prosodic interference and developing tonal encoding sensitivity.

The results appear to support a developmental trajectory in the acquisition of T4 that corresponds with increased academic exposure and phonological refinement. However, several observations warrant further discussion to interpret these findings in greater depth. BA2 students appear to struggle with tone realization under prosodic pressure, exhibiting signs of F0 contour manipulation, especially in question conditions. BA3 learners demonstrate partial mastery but also exhibit variability in contour complexity and slope, especially under question conditions. In contrast, MA1 learners tend to exhibit more stable and canonical falling contours, which may indicate an emerging proficiency in managing the lexical tonal specifications of T4 within broader intonational structures. However, as also observed in the Grade.Sentence Type.Focus model for the T1 subset, subtle contour divergences across sentence conditions in MA1 productions suggest that successful lexical tone realization does not necessarily entail full integration of prosodic information at the utterance level. This nuance invites a more cautious interpretation: while MA1 learners appear to have internalized the phonological shape of T4, particularly its falling trajectory, their pitch contours remain relatively invariant across syntactic and pragmatic contexts. The relative uniformity observed in the MA1 tonal contours across different sentence conditions therefore raises the possibility that these learners are not yet fully deploying T4's prosodic potential, and that intonational modulation remains underdeveloped at this stage of acquisition.

6.6.3.3 Interaction of Proficiency, Sentence Type and Focus

To further examine the role of L2 proficiency in the realization of Mandarin T4 contours, we refitted a GAMM with fREML method incorporating a three-way interaction between Proficiency, Sentence Type, and Focus (PSFT4) with Speaker and OtherTone as random intercepts[31].

The fixed effects of Proficiency level were generally not statistically significant in the parametric component of the model, with the exception of High.statement.post, which yielded

[31] Model diagnostics indicated good convergence (max absolute gradient < 10^{-5}) and an adequate basis dimension for all smooths (k-index ≥ 1, $p > 0.49$), suggesting that the functional form of each trajectory was well captured by the smoothing terms. The model explained approximately 34.3% of the deviance, with an adjusted R^2 of 0.322.

a significant negative coefficient (β = -0.146, p = 0.012). This finding suggests that high proficiency learners may lower their F0 in post-focus statement contexts, possibly reflecting a more native-like declination pattern associated with PFC.

However, the most informative insights derive from the smooth terms, which revealed strong and significant effects of proficiency on the shape and complexity of the F0 contours across conditions. Specifically, high proficiency learners produced more stylized and compressed contours in on-focus questions and statements, as evidenced by significant smooth terms with relatively low estimated degrees of freedom (edf ≈ 1.0-2.4), suggesting more linear tonal trajectories. Low proficiency learners, by contrast, exhibited more complex and variable contours across most conditions, with higher edf values (up to 3.7) and broader confidence intervals. Importantly, GAMM-derived pitch contours reveal that in both question contexts, low-proficiency learners display a late rising trend in the final portion of the syllable, potentially reflecting negative prosodic transfer from their L1. In contrast, high-proficiency learners consistently maintain the expected falling trajectory, indicative of increased tonal stability under prosodic pressure. However, this stability may come at the expense of prosodic flexibility, potentially limiting the integration of intonational cues required for effective discourse marking.

Visualizations of the model further support this interpretation (see Figg. 139-142). Across on-focus contexts, high-proficiency learners demonstrated greater control over pitch trajectory, yielding consistently falling contours, while low-proficiency learners exhibited more irregular and flattened contours indicative of unstable tonal realization. In post-focus conditions, both groups implemented some degree of pitch lowering, although this pattern was most salient in statements, suggesting a developing awareness of PFC.

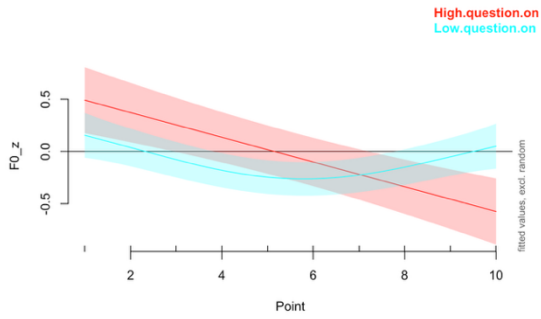


Figure 139 Tone 4 question on-focus by Proficiency

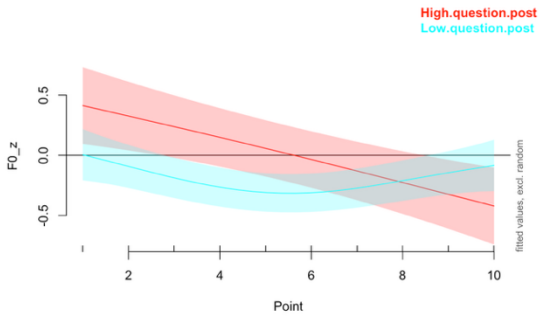


Figure 140 Tone 4 question post-focus by Proficiency

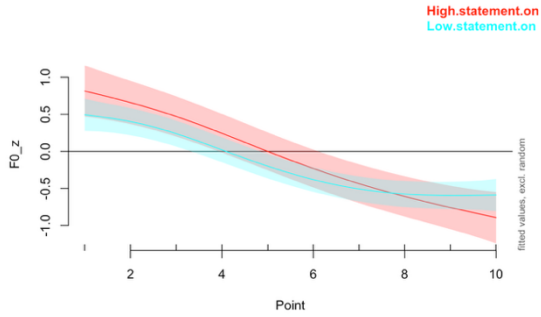


Figure 141 Tone 4 statement on-focus by Proficiency

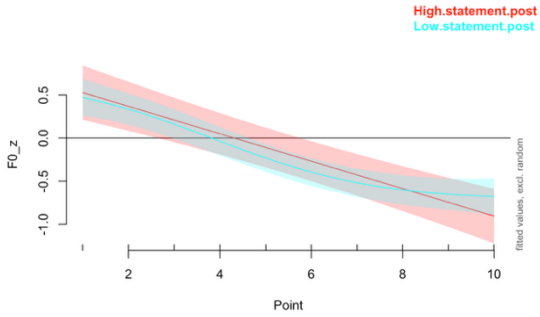


Figure 142 Tone 4 statement post-focus by Proficiency

6.6.4 Interim summary on T4

Native speakers consistently demonstrated dynamic, steeply falling F0 contours characteristic of T4, with robust modulation across pragmatic contexts. Italian learners exhibited shallower slopes, reduced pitch excursion, and attenuated F0 maxima, particularly in question contexts, where native speakers showed pitch enhancement. This contrast was most pronounced in on-focus questions, where the interaction between sentence type and discourse prominence elicited the greatest divergence in pitch trajectory between groups. In fact, in learners' production of question the contour was not strictly falling but rather exhibited a subtle falling-rising pattern. This contour shape suggests a possible prosodic interference from the learners' L1 intonation system, in which final rising pitch is commonly associated with yes-no questions (L-H%; see § 2.1.3). The deviation from the canonical falling shape of T4 may reflect learners' attempts to simultaneously encode lexical tone and sentence-level prosody, albeit with limited success in prosodic integration.

Post-focus conditions – typically associated with pitch compression in Mandarin – highlighted further group differences. While native speakers displayed clear post-focal

lowering in both slope and minimum pitch (F0_min), Italian learners only partially implemented this prosodic feature. Importantly, in statement-post contexts, learners' F0 trajectories approximated native-like pitch height, but retained a high-convex citation-like contour, suggesting persistent reliance on lexical tone templates learned in isolation. This underscores a limited prosodic flexibility and a lack of integration between lexical tone and prosodic structure in L2 productions.

Analyses of F0_range revealed that Italian learners produced narrower pitch excursions, regardless of sentence type or focus structure. Although both groups reduced pitch range in statements compared to questions, native speakers exhibited greater modulation, aligning with Mandarin's tonal-prosodic system. Learners, by contrast, showed compressed range even in interrogatives, pointing to reduced tonal expressivity and prosodic rigidity – factors that may impede communicative effectiveness in tonal languages.

Model comparisons incorporating individual learner variables revealed that grade level was the most robust predictor of T4 performance in these experimental conditions. MA1 learners produced more canonical T4 contours and exhibited improved control over F0 scaling and slope, particularly in prosodically complex environments (e.g., post-focus contexts). However, even advanced learners displayed signs of prosodic inflexibility – maintaining uniform tone contours across contexts that would otherwise demand pitch modulation.

Interestingly, Proficiency as a categorical variable yielded modest improvements in model fit, indicating that general language ability contributes to tonal control, though less strongly than academic progression. In contrast, Musicality, which had been a significant predictor in the realization of rising tones (T2), did not significantly influence T4 production.

Low-proficiency learners frequently exhibited late rising pitch patterns in interrogative contexts, possibly reflecting negative transfer from the intonational structure of Italian. In contrast, high-proficiency learners maintained the expected falling trajectory of T4, reflecting greater tonal stability. However, their reduced contextual flexibility and tendency toward uniform pitch patterns suggest a trade-off between segmental tonal accuracy and prosodic adaptability – a finding consistent with prior studies on L2 tone acquisition (Yang, 2016; Cao 曹文, 2022; Juhász, Bartos, 2022*; inter alia*).

These findings reinforce the notion that accurate tone realization alone is not sufficient for native-like prosodic competence in a tonal language. L2 learners of Mandarin, even at advanced stages, may acquire the phonemic contour of tones like T4, but fail to fully integrate them into sentential prosody. Moreover, flattened pitch contours, attenuated slope, and reduced

F0_range in L2 productions suggest persistent challenges in pitch control, possibly shaped by motoric constraints, negative phonological transfer, or instructional gaps in prosodic training. The asymmetry in learners' treatment of tonal categories further suggests that tonal salience and contour shape interact with cognitive and perceptual factors in L2 tone acquisition.

6.7 Discussion

The study presented in this chapter investigated how Italian university learners of Mandarin Chinese employ tonal contours in phrase-final position to encode sentence type, comparing their productions with those of native Mandarin speakers. Specifically, it examined whether learners disregard lexical tonal specifications in favor of boundary-tone-like prosodic cues – a tendency consistent with their L1 prosodic system – to signal interrogativity, and how such strategies diverge from native patterns. Drawing on a corpus of annotated disyllabic tokens, the analyses focused on three tones: Tone 1 (T1), Tone 2 (T2), and Tone 4 (T4).

The results demonstrate that while Italian learners have largely acquired the phonemic contours of Mandarin tones, they face persistent difficulty in integrating these tones into broader prosodic structures. Native speakers consistently modulated tonal shape and pitch range according to sentence type and focus condition, producing systematic expansions and compressions characteristic of Mandarin's tonal-prosodic system. In contrast, learners frequently defaulted to citation-like contours with limited contextual modulation, yielding flatter trajectories, narrower pitch ranges, and weaker differentiation across pragmatic conditions.

Native speakers realized T1 as a stable high pitch but systematically modulated it – raising F0 in interrogatives and compressing contours post-focus. Italian learners, however, produced lower maxima, shallower trajectories, and weaker differentiation between sentence types. Proficiency strongly predicted L2 performance: high-proficiency learners produced stable high pitch in question-on contexts and context-sensitive adjustments in statements, partially approximating native-like targets. Low-proficiency learners, by contrast, showed greater pitch variability and more irregular contours, suggesting reduced prosodic control. Importantly, even advanced learners retained divergences in interrogatives, with limited contour adjustments compared with native speakers.

T2 offered a particularly diagnostic window into L2 prosodic integration. Native speakers frequently undershot the rising contour in post-focal or declarative contexts, producing compressed or flattened realizations. Learners, however, retained rising slopes across all

conditions, effectively overapplying the citation form. This mismatch was evident in slope, maximum pitch, and pitch range, all of which remained elevated relative to compressed native targets. Sentence-type contrasts were likewise attenuated: whereas native speakers expanded pitch in questions, learners showed reduced differentiation between interrogatives and statements. Within the learner group, Proficiency was less predictive than academic progression and musical aptitude, with musically trained or more advanced students displaying more nuanced control.

For T4, native speakers produced steeply falling contours modulated by sentence type and focus (e.g., interrogative expansion and post-focus compression). Learners diverged most sharply in interrogative contexts: instead of a steep fall, they often produced a falling-rising trajectory – consistent with transfer from learners' L1 yes-no question intonation (L-H%). Even when pitch height approximated native levels (e.g., in statement-post contexts), learners' contours typically retained a high, convex, citation-like shape rather than the compressed native pattern. Grade level was the strongest predictor of performance: MA learners produced more canonical falls but still exhibited prosodic rigidity, suggesting a developmental trade-off between tonal accuracy and contextual flexibility.

Addressing RQ1, the findings confirm that learners primarily manipulate contour shape rather than pitch register, as hypothesized. This tendency is most evident for Tone 4 (T4), where learners frequently substitute rising or convex contours for canonical falls to mark questions, irrespective of focus condition, reflecting alignment with Italian L1 intonational strategies. In contrast, for T1 and T2, learners predominantly rely on citation forms, showing limited adaptation to prosodic or pragmatic context.

Addressing RQ2, the results reveal systematic misalignments between native speakers and learners. Whereas native speakers dynamically integrate tone and intonation – compressing, expanding, or undershooting tones depending on sentence type and focus – Italian learners tend to preserve citation-like tones across contexts. As hypothesized, this pattern indicates reliance on contour-based strategies rooted in L1 prosody rather than target-like modulation of pitch register.

Both Proficiency and Grade (academic progression) emerged as significant predictors, though their influence varied across tones. For T1, Proficiency was decisive: more advanced learners demonstrated context-sensitive adjustments. For T2 and T4, academic progression was the stronger predictor. These results partly confirm the hypothesis that contour-based strategies

persist across proficiency levels while suggesting that advancement through academic stages stabilizes citation forms before prosodic modulation is mastered.

Taken together, these findings underscore that successful acquisition of Mandarin tones by learners from non-tonal L1 backgrounds requires more than phonemic accuracy. Although learners reproduce citation contours with increasing precision, they often fail to implement the prosodic flexibility necessary for integrating tones into sentence-level intonation. This asymmetry reveals a dissociation between lexical tone acquisition and prosodic integration, implying distinct developmental trajectories for lexical-level versus discourse-level features.

The persistent reliance on citation forms and rising contours in interrogatives further suggests cross-linguistic transfer from Italian, where boundary tones and pitch accents function as primary cues to sentence type. Without explicit pedagogical focus on prosody and discourse-level pragmatics, learners are likely to continue relying on L1-based strategies that conflict with Mandarin norms.

Future research should proceed in two directions. First, perception studies are needed to assess whether native listeners interpret L2 productions as pragmatically informative or merely lexically accurate. Such evidence could clarify whether reliance on citation forms hampers communicative effectiveness. Second, longitudinal and pedagogical studies should test how explicit training in prosodic integration – particularly in pitch-range expansion and compression – might accelerate acquisition. Finally, individual-difference variables such as musical aptitude merit further investigation, as they appear to facilitate tonal modulation only partially.

In summary, this study shows that while lower- and upper-intermediate Italian learners of Mandarin successfully acquire the phonological properties of lexical tones in isolation, they encounter enduring difficulties integrating these tones within sentence-level prosody. Native speakers employ systematic adjustments in pitch height, slope, and range to coordinate lexical tones with sentence type and focus, whereas learners often rely on citation-like realizations or exhibit L1-driven prosodic transfer.

A nuanced pattern emerges across tones and learner groups. For T1 and T2 – and for T4 among MA learners – students generally maintained citation-like contours even in contexts where native speakers compressed or modulated tonal trajectories. This strategy preserves lexical meaning but fails to convey prosodic meaning. Conversely, T4 productions among BA students revealed a different pattern: rather than maintaining the falling citation contour, many shifted toward falling-rising trajectories to mark questions, likely reflecting negative transfer

from Italian intonation patterns. These realizations entail both prosodic misalignment and distortion of the lexical tone itself.

On this basis, we may hypothesize that although T4 is, together with T1, among the simplest tones to acquire in isolation, it poses the greatest challenge in intonational contexts. The developmental trajectory observed here suggests that lexical acquisition precedes intonational acquisition. Even when MA students preserved T4's canonical falling contour in phrase-final position, they did not employ prosodic adjustments to convey discourse-level meaning – mirroring the patterns also found for T1 and T2.

Future perceptual work will be essential to test whether, for T1 and T2, lexical tone identity is maintained at the expense of prosodic meaning, whereas for T4 – particularly among less advanced learners – both lexical identity and prosodic function are compromised within the intonational phrase. As the findings of this thesis indicate, the phonological forms of the tones are already well mastered in isolation; the observed misalignments arise not from tonal deficits but from the additional intonational demands of embedding tones within sentence-level prosody. This highlights the pedagogical need to move beyond lexical tone accuracy and explicitly address tone-intonation integration in L2 instruction, particularly for intermediate learners.

7 General discussion and Conclusions

Across the three studies presented in this thesis, a coherent picture emerges of how Italian learners of Mandarin acquire and deploy T1, T2, and T4 under the interacting conditions of focus and sentence type.

Tone 1 (T1), the high-level tone, emerged as the most stable and accurately produced in its citation form. In isolation, learners consistently reproduced its flat contour and appropriate pitch scaling, indicating solid phonological mastery. In disyllabic productions, T1 remained steady in syllable-initial position but exhibited slight variability phrase-finally, suggesting minor boundary-related effects. Under focus, however, only native speakers demonstrated systematic cues for focus marking, such as clear pre-focus lowering and post-focus compression (PFC). Across sentence-type conditions, native speakers modulated T1 through higher onsets and slight contour raising in interrogatives, whereas learners displayed minimal differentiation across contexts. Proficiency emerged as the strongest predictor of T1 performance: higher-proficiency learners maintained more stable high-register control, but their production, while lexically accurate, remained prosodically rigid.

Similarly, Tone 2 (T2), the rising tone, was comparatively stable in isolation, but proved diagnostically revealing in connected speech. In isolated and disyllabic contexts, T2 was accurately shaped but often hyperarticulated by lower-level learners, who over-raised the pitch target regardless of position. In focus conditions, native speakers deepened the rise under focus and reduced it post-focus, whereas learners maintained consistently elevated slopes, showing little sensitivity to discourse context. Even advanced learners failed to implement PFC, producing hyperarticulated rises across focus and sentence-type conditions. This persistent over-specification reflects reliance on phonemic rather than prosodic control: tones are produced "correctly" in form, but not flexibly in function. Sentence-type contrasts further revealed that learners' rising patterns remained uniform across statements and questions, contrasting with native undershooting in declaratives and expansion in interrogatives.

Tone 4 (T4), the falling tone, displayed the most pronounced divergences between native and learner productions in connected speech and followed a distinct developmental trajectory. While it is relatively well mastered in isolation, its integration into broader intonational contours remains fragile. In isolated productions, learners generally performed well, although their realizations showed less consistent onset height and shallower fall slopes than those of native speakers. In disyllabic contexts, proficiency strongly predicted more target-like realizations, with higher-level learners exhibiting earlier pitch lowering and more natural

declination trajectories, particularly when T4 occurred in the second syllable. This positional advantage appears to enhance perceptual salience, thereby facilitating both tonal identification and articulatory planning; it may therefore be hypothesized that T4 is comparatively easier to manipulate within intonational structures. However, under focus conditions, whereas native speakers produced steep on-focus falls followed by clear post-focus compression (PFC), learners frequently flattened the fall or neutralized on/post-focus contrasts altogether. In fact, the most salient negative transfer effects emerged in sentence-type conditions: while native speakers maintained steep falls in statements and used controlled pitch expansion in interrogatives, Italian learners often replaced the canonical fall with a falling-rising or convex contour – an intonationally motivated but phonologically misleading adaptation that mirrors the L-H% boundary tone commonly reported for yes-no questions in several Italian varieties (§ 2.6.1).

This phenomenon leads us to argue that, although T4 is widely recognized in the literature as one of the easiest tones to acquire in isolation, it emerges as the most challenging to produce accurately in its phonological contour under prosodic pressure – particularly in interrogative contexts. In contrast, as outlined above, T1 and T2 appear more prosodically rigid, possibly because their high-level and rising contours afford less flexibility for contextual modulation. Paradoxically, T4 – while easier for learners to master in isolation – proves the most susceptible to tonal contour alteration under sentence-level prosodic demands. Such instability risks compromising both lexical tone identity and pragmatic prosodic meaning, revealing a critical point of fragility at the tone-intonation interface in L2 Mandarin.

Overall, these findings point to a persistent bottleneck at the tone-intonation interface, where learners must reconcile the competing demands of maintaining lexical tone identity while simultaneously encoding sentence-level pragmatic meaning. The recurring substitution of rising contours for T4 in interrogatives illustrates this tension: learners are able to produce the tone accurately in isolation, yet they often fail to embed it within the appropriate prosodic framework, underscoring the incomplete integration between tonal and intonational systems, even among high-proficiency and postgraduate learners of Mandarin.

The present findings thus highlight a systematic gap between native Mandarin speakers and Italian L2 learners in the integration of lexical tone within sentence-level prosody. Learners can approximate the citation forms of Mandarin tones, yet exhibit limited prosodic flexibility, especially in adapting pitch contours to discourse-level functions such as focus and sentence type. This pattern corroborates broader evidence that L2 learners typically achieve lexical

accuracy before mastering the dynamic, context-sensitive modulation that underpins native-like prosodic integration. Within the framework of the L2 Intonation Learning Theory (LILt; Mennen, 2015), the observed difficulties can be then analyzed along four interrelated dimensions: systemic, realizational, semantic, and frequency (see § 1.1.3 for an overview on LILt dimensions).

Systemic Dimension

The systemic dimension concerns the inventory of prosodic categories and their permissible combinations. Italian learners in this study clearly possess the phonological categories of Mandarin tones: baseline analyses confirm that target tones are well established in isolation, with appropriate pitch scaling and contouring. However, systemic divergences emerged once tones were embedded into intonational phrases. For instance, BA learners frequently replaced canonical falling T4 contours with falling-rising shapes in interrogatives, reflecting direct transfer of Italian L-H% boundary tones. This systemic substitution distorts the Mandarin tonal inventory: rather than maintaining T4's lexical fall, learners re-map interrogative marking onto a boundary-tone-like contour, producing hybrid tonal-intonational configurations absent from native Mandarin.

Realizational Dimension

Realizational mismatches proved the most pervasive. Native Mandarin speakers modulated tones dynamically – sustaining stable high T1 contours with elevated F0 in questions and lowered post-focus; executing steep, focused rises and compressed post-focus realizations for T2; and producing steep, context-sensitive T4 falls with strong post-focus lowering. Italian learners, by contrast, defaulted to citation-like contours: T1 was shallower and exhibited downward drift; T2 maintained rising slopes across all contexts without sufficient PFC and undershoot; and T4 falls were reduced in both slope and range. Learners also undershot pitch minima (F0_min) across tones, especially in T2 and T3, restricting the overall pitch span – an issue consistent with earlier observations of low-register avoidance in L2 intonation (Mennen, 2004). Even advanced learners (MA1) who produced canonical T4 falls often failed to modulate them across contexts, reflecting realizational rigidity rather than phonetic inaccuracy. Musical aptitude facilitated finer control of T2's rising slope and post-focus flattening but had minimal effects on T4, suggesting tone-specific constraints in auditory-motor tracking.

Semantic Dimension

At the semantic level, learners' deviations impaired the communicative effectiveness of prosody. Native Mandarin speakers used on-focus enhancement combined with robust PFC to mark information structure, highlighting focused constituents while deaccenting post-focal material. Learners instead relied on global F0 raising without corresponding post-focus lowering, leading to weak on/post-focus contrasts. For T2, learners produced consistently higher rises in all contexts rather than modulating them for statement vs. question contrasts, thereby blunting pragmatic differentiation. In interrogative type, the use of falling-rising T4 violated Mandarin prosodic norms and potentially compromised the prosodic information and the lexical tone itself. Hence, while T1 and T2 maintained lexical identity at the cost of prosodic nuance, T4 often entailed a dual distortion – compromising both lexical and prosodic meaning. These asymmetries underscore how intonational pressure can distort lexical tone identity, particularly for low-target tones such as T4.

Frequency Dimension

Frequency patterns reveal entrenched reliance on L1-based intonational defaults. Italian learners systematically overused rising contours across contexts, particularly for T2, where citation-like rises persisted even in post-focus environments that required compression. Similarly, T4 interrogatives frequently contained rising tails, directly mirroring Italian yes-no question intonation. This over-reliance on rising contours echoes cross-linguistic findings in L2 prosody (see § 1.1.3 and references therein), where learners generalize rises as default cues for prominence and interrogativity regardless of target-language constraints.

Table 45 Summary of cross-dimensional differences between native Mandarin speakers and Italian learners according to the L2 Intonation Learning Theory (LILt) framework

LILt Dimension	Native Mandarin Patterns	Italian Learner Patterns	Tone-specific Findings
Systemic	Phonologically stable tonal inventory integrated with intonation	Hybrid categories via L1 transfer (e.g., T4 falling-rising contours in interrogatives)	T4 (BA learners): systemic substitution of Italian L-H% boundary tone
Realizational	Context-sensitive modulation: T1 high and flat with elevation in questions, T2 steepened rises + PFC, T4 steep falls + PFC	Citation-like contours, undershot minima, compressed ranges, weak modulation	T1: downward drift. T2: invariant rise. T4: shallow or rising tail.
Semantic	Focus marked by on-focus enhancement + PFC;	Focus marked by global F0 raising;	T1/T2: lexical preserved, prosody

LILt Dimension	Native Mandarin Patterns	Italian Learner Patterns	Tone-specific Findings
	interrogativity by expanded pitch range	interrogativity by rising/convex shapes	flattened; T4: lexical + prosody both compromised
Frequency	Balanced use of rises/falls across tones and contexts	Over-reliance on rises, reflecting L1 bias	T2: rising maintained universally; T4: rising tails frequent in questions

Overall, the findings indicate that intermediate Italian learners struggle less with acquiring tonal categories than with integrating them into Mandarin's prosodic framework. The evidence supports a developmental trajectory in which lexical tone acquisition precedes prosodic flexibility, yet persistent L1-driven realizational, semantic, and frequency mismatches constrain learners' ability to convey pragmatic meaning effectively.

Generalizability of the Findings to Sentence-final Particle Questions

As specified earlier, this study exclusively examined unmarked echo questions. However, for the sake of ecological validity, it is important to note that Mandarin interrogatives are frequently marked syntactically by means of modal particles, which are themselves neutral-tone syllables. The generalizability of the present findings to intonational phrases ending with such particles – most notably ma 吗, ba 吧, and ne 呢 – is therefore particularly relevant. Although this issue requires verification through targeted control studies, some cautious preliminary considerations can be advanced here.

We argue that such generalization is plausible, given that neutral-tone syllables are highly susceptible to intonational coarticulation[32] (see § 2.5.2). In this view, the final lexical tone preceding the particle can be regarded as carrying the intonational nucleus, whereas the phrase-final neutral particle functions as a peripheral prosodic element appended to the intonational phrase. Consequently, the pitch realization of the particle would be conditioned by the boundary configuration of the preceding tone.

For instance, in a L2 Mandarin disyllabic T4-T4 sequence that realizes a boundary-tone-like LH% rise in phrase-final position, the rising trajectory may extend into the onset of the neutral particle, thereby spreading beyond the tone-bearing lexical domain.

[32] However, it should also be acknowledged that the prosodic cues observed in the present study – elicited through unmarked echo questions – may be more salient than those found in syntactically marked interrogatives, as is generally reported for L1 Mandarin prosody (§ 2.6.1).

Such coextension of pitch movement supports the hypothesis that, in L2 Mandarin speech, where prosodic phrasing and tonal anchoring are less tightly integrated, boundary tones can override lexical tonal identity, resulting in perceptual and articulatory overlap between lexical and intonational functions. Extending this analysis to ma-marked and unmarked interrogatives will thus be essential to determine whether the same asymmetries observed in this study persist under variable boundary conditions (see § 7.2).

7.1 Implications for second language research and teaching

As outlined in § 1 and § 2, prosody is a fundamental component of speech organization, interpretation, and communicative effectiveness. Yet, research on second language (L2) acquisition has traditionally prioritized segmental accuracy, producing highly influential models such as the Speech Learning Model (Flege, 1995) and the Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007). Although these models have advanced our understanding of segmental learning, they offer limited explanatory scope for suprasegmental phenomena, where meaning is distributed across prosodic structure rather than individual phonetic segments. The present findings demonstrate that segmental and lexical tone accuracy can precede – and even dissociate from – prosodic integration, confirming that suprasegmental acquisition requires dedicated theoretical and practical treatment.

The results provide strong empirical support for the L2 Intonation Learning Theory (LILt; Mennen, 2015), which differentiates four dimensions of divergence in L2 prosody: systemic, realizational, semantic, and frequency. Italian learners of Mandarin in this study showed that while systemic tone categories were successfully acquired, persistent realizational mismatches (e.g., undershot pitch floors, invariant rises, shallow falls) and semantic miscuing (e.g., global F0 raising without post-focus compression) remained evident even among advanced learners. These results confirm that L2 intonational development entails the restructuring of entrenched L1 prosodic routines, not merely the addition of new phonological categories.

Mandarin provides a particularly stringent test case for models of suprasegmental acquisition, as the same acoustic dimension – fundamental frequency – simultaneously encodes lexical tone and intonation (§ 2.5.2). Analyses of Tones 1, 2, and 4 indicate that learners from non-tonal L1 backgrounds are generally able to produce citation tones in isolation; however, they struggle to deploy these tones flexibly in response to discourse-driven factors such as focus or sentence type. This dissociation highlights the importance of conceptualizing tones not as fixed, invariant targets but as dynamic phonological resources whose realization is

conditioned by higher-level prosodic structure within intonational phrases. Promoting this functional perspective on Mandarin tones should occur not only at advanced stages of acquisition – when learners have already internalized tones as static categories – but critically at early stages, introducing prosodic phenomena through minimal intonational phrases. While pedagogical activities informed by this perspective are proposed below, further applied research is needed to empirically validate the approach (see § 7.2).

These findings align with current frameworks such as the Common European Framework of Reference (CEFR, 2018), which prioritize intelligibility and comprehensibility (Piccardo & North, 2017). Italian learners' apparent tonal accuracy often conceals deficits in pragmatic intelligibility: they may produce lexically correct words that fail to signal focus or sentence type, thereby compromising discourse-level understanding. Traditional curricula – especially in European contexts – tend to emphasize tone drills that isolate tone citation forms without sufficient integration into discourse-level tasks. The present findings including upper-intermediate learners demonstrate that this approach is insufficient for achieving communicative adequacy in L2 Mandarin.

To address these shortcomings, instruction should target prosodic flexibility and context-sensitive pitch control at least through the following principles:

- Post-focus compression (PFC);
- Pitch-range management;
- Tone-specific modulation;
- Contrastive and communicative practice: embedding tones in dialogues that require alternation between statements and questions or between new and given information.

In line with recent pedagogical reforms, the goal of L2 prosody instruction should shift from accent elimination to functional proficiency. For Mandarin, this entails ensuring that tones are both lexically accurate and prosodically meaningful across a range of communicative contexts and speaker varieties. Crucially, learners should be exposed not only to Standard Mandarin Chinese but also to regional and sociolectal varieties that differ in pitch range, tonal implementation, and intonational contouring. Such exposure reflects the sociolinguistic reality of contemporary Mandarin use and prevents learners from equating "correctness" with a single standardized accent. Moreover, awareness of legitimate variation within the Mandarin phonological space may enhance learners' motivation and foster greater communicative adaptability. Familiarity with tonal and prosodic variability – such as the flatter pitch patterns

typical of Beijing Mandarin or the wider tonal excursions of southern accents – can enhance perceptual robustness and promote adaptive control of F0 in production. In this sense, the inclusion of non-standard pronunciation models in pedagogical materials, such as that proposed by Raini and Wang (2023), represents an important advance: the deliberate incorporation of regional accent features provides learners with a more authentic range of input and reinforces the development of flexible, context-sensitive prosodic competence. Broadening exposure beyond a single prestige norm thus supports both communicative flexibility and sociophonetic awareness, enabling learners to interpret and participate effectively across diverse Mandarin-speaking contexts.

Furthermore, shadowing-based activities may prove crucial in light of these results. Shadowing – an online, real-time repetition technique – has been shown to enhance L2 pronunciation by improving both perceptual acuity and phonological memory (Kadota, 2007; Hamada, 2015; Yang, 2019), and it may also be particularly effective for Mandarin pronunciation acquisition (Francolino, 2022; 2024b; Raini & Wang, 2023). Through the temporary retention and reactivation of auditory traces in short-term memory, learners progressively consolidate them in long-term memory, supporting the formation of a stable phonetic-lexical repertoire (Nye & Fowler, 2003). As Kadota (2019) argues, sustained shadowing fosters increasingly automatic and unconscious assimilation of linguistic input, thereby reducing cognitive load and facilitating the internalization of prosodic and tonal patterns. Integrating shadowing into L2 pedagogy thus offers both sensorimotor and cognitive reinforcement, consolidating the flexible tonal control essential for fluent, context-appropriate Mandarin speech. Particularly relevant in this regard are several studies on the acquisition of Italian intonation as a foreign language among Chinese learners, which have demonstrated that self-imitative techniques are more effective than traditional imitation exercises in enhancing intonational features in L2 speech (De Meo et al., 2013; De Meo et al., 2016; Vigliano et al., 2016). Extending this line of inquiry to Italian learners of Mandarin could yield valuable insights, as similar self-monitoring and auto-imitative mechanisms – among which shadowing itself constitutes a particularly effective technique – may facilitate the development of prosodic awareness and control within tonal-intonational systems.

By coupling tone practice with functional discourse tasks and exposure to multiple varieties of Mandarin, learners can begin to conceptualize tones as adaptive phonological units rather than fixed pitch templates. This orientation fosters both lexical precision and intonational appropriateness, leading to greater overall communicative effectiveness.

In sum, the evidence presented in this thesis underscores that prosodic competence is a core component of L2 communicative adequacy. Although Italian learners of Mandarin achieve reliable accuracy in isolated tone production, their difficulty in integrating tones into intonational structures undermines both intelligibility and comprehensibility in discourse. Pedagogically, these results call for a reorientation of teaching practice – away from static tone imitation and toward dynamic, context-sensitive prosody training that reflects the full range of Mandarin phonological diversity. In fact, the central challenge in L2 Mandarin learning may not be the articulation of tones themselves, but their functional deployment as communicative resources within the tonal-intonational system.

7.2 Limitations and future directions

This study presents several methodological and conceptual limitations that, while acknowledged, remain only partially addressed within the current design. Future research will need to investigate these issues in greater depth to substantiate the claims advanced here and to broaden the scope of L2 prosodic inquiry.

Methodological Constraints

A primary limitation concerns the use of a scripted reading task, which – although advantageous for maintaining strict control over variables such as character recognition, lexical retrieval, and syntactic planning – may not fully capture the dynamic properties of spontaneous speech. The inclusion of short, scripted dialogues with disyllabic target phrases was intended to elicit tonal productions under controlled prosodic conditions and to heighten learners' awareness of contrasts between statement and echo-question contexts. However, this design also introduces the possibility of prosodic accommodation, whereby the intonation of a question may be influenced by the preceding statement. While this effect does not undermine the main findings, it suggests that complementary tasks – such as interactive board games, role-play scenarios, or guided interviews – would be valuable in eliciting a broader range of sentence types and focus structures under more naturalistic, though less controlled, conditions.

A second limitation lies in the restricted segmental variability of the stimuli. Each target phrase was represented by a single segmental configuration, limiting the generalizability of the results and precluding systematic analysis of how phonological environment interacts with tonal realization. Although random intercepts in statistical modeling partially mitigate item-level effects, they cannot substitute for a more balanced design incorporating diverse segmental

contexts. Expanding future datasets to include a wider range of phonetic environments will allow for more robust testing of how segmental-suprasegmental interactions shape L2 tonal prosody.

In fact, the methodological trade-off underlying this project reflects a broader challenge in L2 phonological research: balancing experimental control with ecological validity. Reading tasks afford precision in isolating tonal and prosodic variables but inevitably constrain the natural variability of learner speech. Conversely, interactive tasks yield more authentic data but introduce variability that complicates modeling and cross-condition comparison (see Xu, 2010 vs. Wagner et al., 2015). A productive future direction lies in combining these approaches – using controlled stimuli to establish baselines while also incorporating semi-spontaneous interactional tasks to test whether prosodic patterns persist in authentic communicative settings. Such a dual-method approach would strengthen both the reliability and ecological validity of findings and illuminate how learners deploy prosody dynamically in real-time conversation.

Furthermore, while the present thesis focuses primarily on quantitative measures, the qualitative data from the post-test questionnaire – though not analyzed in the current work – will be employed in future follow-up studies involving integrated quali-quantitative analyses.

Individual and Sociophonetic Factors

Another limitation pertains to the limited consideration of participants' individual and sociophonetic variability. While this thesis identified significant contributions of proficiency, academic progression, and musical aptitude, it did not directly test how these factors interact with perceptual outcomes. Future work should include perception-based evaluations to determine whether learners' productions are not only phonetically accurate but also pragmatically interpretable to native listeners. This would clarify whether reliance on citation-like forms undermines communicative adequacy and whether tonal distortions – particularly for Tone 4 – are perceived as intelligible or misleading.

A further avenue for refinement involves cross-linguistic transfer. To assess potential L1 prosodic influence more precisely, comparative analyses of learners' L1 varieties are required. However, as no single variety of Italian can be considered a standard beyond a theoretical construct, and given the ongoing debate surrounding Italian prosodic typology (Berruto, 2007; Crocco, 2017), regional variation must be treated as an integral factor. In this study, participants' linguistic backgrounds were controlled through provenance (Rome vs. Siena), but this variable could not be fully incorporated into the statistical analysis due to resource constraints.

Considering that Roman and Sienese speech differ systematically in their intonational profiles – rising-falling for Rome and falling-rising (H+L* L-H%) for Siena (Marotta & Sorianello, 1999), at least in read-speech (see § 2.6.1) – a comparative treatment of these subgroups would help explain the Tone 4 boundary-tone-like rise observed in learners' interrogatives. Nevertheless, in light of increasing bilingualism and population mobility in Italy, the ecological validity of sharply localized prosodic comparisons must also be critically evaluated.

Conceptual and Theoretical Extensions

Beyond methodological considerations, this study opens several conceptual and applied research avenues. First, perceptual validation is essential: whether native Mandarin listeners interpret L2 productions as pragmatically coherent or merely lexically accurate remains an open question. Second, longitudinal studies are needed to test whether prosodic flexibility emerges naturally with exposure or requires explicit instruction – for example, through targeted training in post-focus compression, pitch-floor control, and sentence-type contrast. Such work would bridge the gap between lexical tone mastery and the discourse-level competence necessary for native-like communicative effectiveness in L2 Mandarin.

Third, the role of individual differences warrants further exploration. Musical aptitude, for instance, was found to selectively benefit rising tones (T2) but not falling tones (T4), suggesting that tone-specific perceptual-motor constraints influence acquisition trajectories. Integrating measures of working memory, auditory discrimination, and rhythmic entrainment could help clarify how domain-general cognitive abilities support tone-prosody integration.

Finally, future research should incorporate perception-production links more explicitly, testing whether training interventions that combine auditory exposure, visual pitch feedback, and shadowing-based repetition facilitate the automatization of tonal control. The inclusion of these training interventions in future experimental and classroom-based research would allow systematic testing of how sensorimotor rehearsal and memory consolidation contribute to tone-intonation integration.

At a theoretical level, these lines of inquiry align with Pelzl and colleagues (2019, 2021), who argue that the persistent difficulty of tone acquisition among advanced learners does not primarily stem from perceptual limitations, but from the cognitive demands of integrating lexical tones into multisyllabic and discourse-level representations. Extending this perspective to prosodic integration, future studies should examine how tonal realization interacts with

information structure, syntactic phrasing, and sentence type, and how these mappings evolve with increased exposure and cognitive automatization.

Interim summary

In sum, the current findings provide a robust foundation for future work at the intersection of L2 phonology, prosody, and cognition. By combining controlled experimental design, ecologically valid discourse contexts, and perception-production training paradigms, subsequent research can deepen our understanding of how learners internalize and deploy tone-intonation structures, ultimately informing both theoretical models of suprasegmental acquisition and pedagogical practices aimed at communicative fluency in L2 Mandarin.

References

Adams, C., & Munro, R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterances of some native and nonnative speakers of English. *Phonetica, 35*(2), 125-156.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*(4), 529-555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Ao, B. X. (1993). *Phonetics and phonology of Nantong Chinese* [Doctoral dissertation, The Ohio State University]. OhioLINK ETD Center. https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?p10_accession_num=osu1105384417

Arcodia, G. F., & Basciano, B. (2016). *Linguistica cinese*. Bologna: Pàtron Editore.

Arvaniti, A. (2022). *The autosegmental-metrical model of intonational phonology.* In J. Barnes & S. Shattuck-Hufnagel (Eds.), *Prosodic theory and practice* (Special Collection: CogNet). MIT Press. https://doi.org/10.7551/mitpress/10413.003.0004

Arvaniti, A. (2020). *The phonetics of prosody*. In *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.411

Arvaniti, A., & Ladd, D. R. (1995). Proceedings of the XIIIth International Congress of Phonetic Sciences: ICPhS 95, Stockholm, Sweden (Vol. 4, pp. 420-423).

Athanasopoulou, A., Vogel, I., Han, C., & Yuan, Y. (2019). Confusability of Mandarin Tone 3 and Tone 4: Effects of focus and syllable position. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)* (pp. 491-495).

Austin, L. (1962). *How to do things with words*. Oxford: Oxford University Press.

Avesani, C., Bocci, G., Vayra, M., & Zappoli, A. (2015). *Prosody and information status in Italian and German L2 intonation*. Milano: Franco Angeli.

Backman, N. E. (1979). Intonation errors in second language pronunciation of eight Spanish-speaking adults learning English. *Interlanguage Studies Bulletin, 4*(2), 239-266.

Bao, Z. (1999). *The structure of tone*. Oxford: Oxford University Press.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology* (1st ed., pp. 9-54). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0002

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255-309.

Berruto, G. (2007). Miserie e grandezze dello standard. La nozione di standard, non standard, substandard in linguistica e sociolinguistica. In Molinelli P. (ed.), *Standard e non standard tra scelta e norma* (pp. 13-41). Roma: Il Calamo.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171-232). Timonium, MD: York Press.

Best, C. T., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

Bianco, R., Ptascynski, L. E., & Chait, M. (2020). Long-term implicit memory for sequential auditory patterns in humans. *eLife, 9*, e56073. https://doi.org/10.7554/eLife.56073

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. Journal of Cognitive Neuroscience, 23(2), 425-434.

Boersma, P., & Weenink, D. (2025). *Praat: Doing phonetics by computer* [Computer program] (Version 6.4.43). Retrieved September 14, 2025, from https://praat.org

Braun, B., & Geiselmann, S. (2011). Italian in the no-man's land between stress-timing and syllable-timing? Speakers are more stress-timed than listeners. In *Proceedings of INTERSPEECH 2011: 12th Annual Conference of the International Speech Communication Association* (pp. 689–692). Florence, Italy.

Brugnoli, A. (2024). *Primary and secondary stress in Italian and German: Position, weight-sensitivity and acoustic correlates* (Doctoral dissertation, University of Verona). https://hdl.handle.net/11562/1129726

Cao, M., Pavlik, P. I., Jr., & Bidelman, G. M. (2024). Enhancing lexical tone learning for second language speakers: Effects of acoustic properties in Mandarin tone perception. *Frontiers in Psychology, 15*, 1403816. https://doi.org/10.3389/fpsyg.2024.1403816

Cao, J. 曹剑芬. (1999). Hanyu jiezou de shengxue yuyinxue tezheng [The acoustical features of Chinese rhythm]. In J. Cao (Ed.), *Xiandai yuyinxue lunwen ji* 现代语音学论文集 [Collections of modern phonetic studies]. Jincheng Chubanshe.

Cao, J. 曹剑芬. (2002). Hanyu shengdiao yu yudiao de guanxi 汉语声调与语调的关系 [The relationship between Chinese tones and intonation]. *Zhongguo Yuwen* [Chinese Language], 3, 195-202.

Cao, W. 曹文. (2010). *Han yu jiao dian zhong yin de yun lü shi xian: Pu tong hua tong wen yi jiao ju de shi yan yan jiu* 汉语焦点重音的韵律实现：普通话同文本异焦句的实验研究 [The prosodic realization of focus accent in Mandarin Chinese: An experimental study on contrastive focus sentences with identical texts]. Beijing: Beijing Yuyan Daxue Chubanshe.

Cao, W. 曹文. (2022). *Yǔyīn jí yǔyīn xídé yánjiū* 语音及语音习得研究 [Chinese phonetics and its acquisition]. Beijing: Beijing Language and Culture University Press.

Casentini, M., & Francolino, D. (unpublished). *Uncovering the strength of weak elements in a tone language: A phonetic study of two neutral-tone sentence-final particles in Mandarin Chinese*. Presentation delivered at the AISV Annual Conference *The Sound of Grammar: New Perspectives on the Interplay of Phonetics with Morphology, Syntax and the Lexicon*, Università degli Studi di Urbino Carlo Bo, February 2025.

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). A system of tone letters. *Le Maître Phonétique, 45*, 24-27.

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology*, Academia Sinica. https://doi.org/10.6355/BIHPAS.193301.0121

Chao, Y. R. (1968). *A grammar of spoken Chinese* (2nd print). Berkeley: University of California Press.

Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America, 126*(3), 1394-1405. https://doi.org/10.1121/1.3177275

Cheang, H. S., & Pell, M. D. (2011). Recognizing sarcasm without language: A cross-linguistic study of English and Cantonese. *Pragmatics & Cognition, 19*(2), 203-223. https://doi.org/10.1075/pc.19.2.02che

Chen, L., & Li, S. (2023). An experimental study of Chinese disyllabic tone errors in native French speakers and an investigation of online teaching strategies. *SHS Web of Conferences, 174*, 01015. https://doi.org/10.1051/shsconf/202317401015

Chen, L., Li, X., & Yang, Y. (2012). Focus, newness and their combination: Processing of information structure in discourse. *PLOS ONE, 7*(8), e42533. https://doi.org/10.1371/journal.pone.0042533

Chen, L., Wang, L., & Yang, Y. (2014). Distinguish between focus and newness: An ERP study. *Journal of Neurolinguistics, 31*, 28-41. https://doi.org/10.1016/j.jneuroling.2014.06.002

Chen, M. Y. (2000). *Tone sandhi: Patterns across Chinese dialects* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511486364

Chen, Q. (1997). Toward a sequential approach for tonal error analysis. *Journal of the Chinese Language Teachers Association, 32,* 21-39.

Chen, S., Zhang, C., McCollum, A. G., & Wayland, R. (2017). Statistical modelling of phonetic and phonologised perturbation effects in tonal and non-tonal languages. *Speech Communication, 88*, 17-38. https://doi.org/10.1016/j.specom.2017.01.006

Chen, Y. (2006). Durational adjustment under corrective focus in Standard Chinese. *Journal of Phonetics, 34*(2), 176-201. https://doi.org/10.1016/j.wocn.2005.05.002

Chen, Y. (2009). Prosody and information structure mapping: Evidence from Shanghai Chinese. *Chinese Journal of Phonetics, 2*, 123-133.

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

Chen, Y. (2022). Tone and intonation. In C.-R. Huang, Y.-H. Lin, I.-H. Chen, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics*. Cambridge: Cambridge University Press.

Chen, Y., & Braun, B. (2006). Prosodic realization of information structure categories in Standard Chinese. In *Proceedings of Speech Prosody 2006* (Paper 051-0). https://doi.org/10.21437/SpeechProsody.2006-92

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics, 36*(4), 724-746. https://doi.org/10.1016/j.wocn.2008.06.003

Chen, Y., & He, A. W. (2001). Dui bu dui as a pragmatic marker: Evidence from Chinese classroom discourse. *Journal of Pragmatics, 33*(9), 1441-1465. https://doi.org/10.1016/S0378-2166(00)00084-9

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech: Evidence from F_0 patterns of neutral tone in Standard Chinese. *Phonetica, 63*(1), 47-75. https://doi.org/10.1159/000091406

Chen, Y., Xu, Y., & Guion-Anderson, S. (2015). Prosodic realization of focus in bilingual production of Southern Min and Mandarin. *Phonetica, 71*(4), 249-270. https://doi.org/10.1159/000371891

Cheng, C., & Xu, Y. (2015). Mechanism of disyllabic tonal reduction in Taiwan Mandarin. *Language and Speech, 58*(3), 281-314. https://doi.org/10.1177/0023830914543286

Cheng, C.-C. (1973). *A synchronic phonology of Mandarin Chinese*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110866407

Chinese Academy of Social Sciences 中国社会科学院. (2021). *Xiàndài Hànyǔ Chángyòng Cíbiǎo 现代汉语常用词表 [List of commonly used words in Modern Chinese]*. Beijing: Commercial Press.

Chu, M., & Qian, Y. (2001). Locating boundaries for prosodic constituents in unrestricted Mandarin texts. *International Journal of Computational Linguistics & Chinese Language Processing, 6*(1), 61-82.

Connell, B. (2001, May 18). Downdrift, downstep, and declination. In *TAPS Proceedings: Typology of African Prosodic Systems Workshop*, Bielefeld University, Germany.

Connell, B., & Ladd, D. R. (1990). Aspects of pitch realisation in Yoruba. *Phonology, 7*(1), 1-29. https://doi.org/10.1017/S095267570000110X

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume with new descriptors*. Strasbourg: Council of Europe.

Crocco, C. (2017). Everyone has an accent. Standard Italian and regional pronunciation. In M. Cerruti, C. Crocco, & S. Marzo (Eds.). *Towards a new standard. Theoretical and empirical studies on the restandardization of Italian* (pp. 89-117). Berlin/New York: De Gruyter Mouton.

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6th ed.). Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781444302776

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

D'Imperio, M. (2001). *Focus and tonal structure in Neapolitan Italian. Speech Communication, 33*(4), 339-356.

D'Imperio, M. (2002). Italian intonation: An overview and some questions. *Probus, 14*(1), 37-49. https://doi.org/10.1515/prbs.2002.004

D'Imperio, M., & Rosenthall, S. (1999). *Phonetics and phonology of main stress in Italian. Phonology, 16*(1), 1-28.

Dai, J. X.-L. (1998). Syntactic, phonological, and morphological words in Chinese. In J. L. Packard (Ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese* (pp. 103-134). Berlin: Mouton de Gruyter.

De Meo, A., Vitale, M., Pettorino, M., Cutugno, F., & Origlia, A. (2013). Imitation/self-imitation in computer-assisted prosody training for Chinese learners of L2 Italian. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference (PSLLT)* (pp. 90-100). Iowa State University.

De Meo, A., Vitale, M., & Pellegrino, E. (2016). Tecnologia della voce e miglioramento della pronuncia in una L2: Imitazione e autoimitazione a confronto. Uno studio su cinesi apprendenti di italiano L2. In F. Bianchi & L. Leone (Eds.), *Linguaggio e apprendimento linguistico: Metodi e strumenti tecnologici* (*Studi AItLA*, 4) (pp. 13-25). Milano.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-397. https://doi.org/10.2307/3588486

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121-144. https://doi.org/10.1207/s15326969eco0102_2

Duanmu, S. (1994). Against contour tone units. *Linguistic Inquiry, 25*(4), 555-608.

Duanmu, S. (2007). *The phonology of Standard Chinese* (2nd ed., revised). Oxford: Oxford University Press.

Duanmu, S. (2022). Evidence for stress and metrical structure in Chinese. In C.-R. Huang, Y.-H. Lin, & I.-H. Chen (Eds.), *The Cambridge handbook of Chinese linguistics* (1st ed., pp. 361-382). Cambridge: Cambridge University Press.

Eriksson, A., Bertinetto, P.M., Heldner, M., Nodari, R., Lenoci, G. (2016) The Acoustics of Lexical Stress in Italian as a Function of Stress Level and Speaking Style. Proc. Interspeech 2016, 1059-1063.

Feng 冯胜利, S. (1996). 论汉语的"韵律词" *Lùn hànyǔ de "yùnlǜcí"* [On prosodic words in Chinese]. *中国社会科学 Zhōngguó Shèhuì Kēxué* [*Social Sciences in China*], 1, 161-176.

Feng, S. (1998). Prosodic structure and compound words in Classical Chinese. In J. L. Packard (Ed.), *New approaches to Chinese word formation* (pp. 197-260). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110809084.197

Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In H. Papousek, U. Jürgens, & M. Papousek (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches* (pp. 43-61). Cambridge: Cambridge University Press.

Fernald, A., & Kuhl, P. K. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10*(3), 279-293. https://doi.org/10.1016/0163-6383(87)90017-8

Féry, C. (2016). *Intonation and prosodic structure* (1st ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139022064

Féry, C., & Ishihara, S. (Eds.). (2016). *The Oxford handbook of information structure* (1st ed.). Oxford: Oxford University Press.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Timonium, MD: York Press.

Francolino, D. (2022). Acquisizione prosodica del cinese LS: Lo shadowing come proposta didattica. *LEND – Lingua e Nuova Didattica, 3*, 54-70.

Francolino, D. (2024a). Tono e intonazione in cinese LS: Un'analisi preliminare. *SILTA (Studi Italiani di Linguistica Teorica e Applicata), 53*(1), 154-171. ISSN: 0390-6809.

Francolino, D. (2024b). Shadowing. In A. Scibetta (Ed.), *Tecniche didattiche per la lingua cinese: Proposte operative per la scuola secondaria di II grado e per l'università*. Torino: UTET Università.

Francolino, D., & Cao, W. (2024). Prosodic challenges among Italian speakers in L2 Mandarin: Preliminary evidence from falling tone production. In *Proceedings of the 5th International Symposium on Applied Phonetics (ISAPh 2024)* (pp. 22-26). https://doi.org/10.21437/ISAPh.2024-5

Gårding, E., Zhang, J., & Svantesson, J.-O. (1983). A generative model for tone and intonation in Standard Chinese. *Working Papers (Lund University, Department of Linguistics), 25*, 53-65.

Giordano, R. (2006). *The intonation of polar questions in two central varieties of Italian.* In *Proceedings of the 3rd International Conference on Speech Prosody* (Dresden, Germany, May 2–5, 2006).

Goldsmith, J. (1976). *Autosegmental phonology* (Doctoral dissertation). Massachusetts Institute of Technology.

Grabe, E. (2004). Intonational variation in urban dialects of English spoken in the British Isles. In P. Gilles & J. Peters (Eds.), *Regional variation in intonation* (pp. 9-31). Tübingen: Niemeyer.

Greif, M. (2010). Contrastive focus in Mandarin Chinese. *Speech Prosody 2010*.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377-388. https://doi.org/10.2307/2182440

Grice, M., D'Imperio, M., Savino, M., & Avesani, C. (2005). Strategies for intonation labelling across varieties of Italian. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 362-389). Oxford: Oxford University Press.

Gu, W., & Lee, T. (2009). Effects of tone and emphatic focus on F0 contours of Cantonese speech: A comparison with Standard Chinese. *Chinese Journal of Phonetics*, 133-147.

Gu, W., Zhang, T., & Fujisaki, H. (2011). Prosodic analysis and perception of Mandarin utterances conveying attitudes. In *Proceedings of Interspeech 2011* (pp. 1069-1072). https://doi.org/10.21437/Interspeech.2011-402

Gussenhoven, C. (2004). *The phonology of tone and intonation* (1st ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511616983

Gussenhoven, C. (2016). Foundations of intonational meaning: Anatomical and physiological factors. *Topics in Cognitive Science, 8*(2), 425-434. https://doi.org/10.1111/tops.12197

Gussenhoven, C., Chen, Y., Frota, S., & Prieto, P. (2013). Intonation. In *Oxford Bibliographies in Linguistics*. https://doi.org/10.1093/OBO/9780199772810-0072

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223. https://doi.org/10.2307/3588378

Hamada, Y. (2015). Uncovering shadowing as an EFL teaching technique for listening: Learners' perceptions, self-confidence, and motivation. *Annual Research Report on General Education, 17*, 9-22.

Handel, Z. (2014). Historical phonology of Chinese. In C.-T. J. Huang, Y.-H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (1st ed., pp. 576-598). Wiley. https://doi.org/10.1002/9781118584552.ch22

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics, 40*(2), 269-279. https://doi.org/10.1016/j.wocn.2011.11.001

Hayes, B. (1995). *Metrical stress theory: Principles and case studies.* University of Chicago Press.

Heffner, C. C., & Slevc, L. R. (2015). Prosodic structure as a domain-general cognitive framework: A bridge across linguistic and musical rhythm. *Frontiers in Psychology, 6*.

Hepper, P. G., & Shahidullah, B. S. (1994). Development of fetal hearing. *Archives of Disease in Childhood, 71*(2), F81-F87. https://doi.org/10.1136/fn.71.2.F81

Hermes, A., Mücke, D., & Grice, M. (2013). Gestural coordination of Italian word-initial clusters: The case of "impure s". *Phonology, 30*(1), 1-25. https://doi.org/10.1017/S095267571300002X

Hewings, M. (1995). The English intonation of native speakers and Indonesian learners: A comparative study. *RELC Journal, 26*(1), 27-46. https://doi.org/10.1177/003368829502600102

Himmelmann, N. P., & Ladd, D. R. (2008). Prosodic description: An introduction for fieldworkers. *Language Documentation & Conservation, 2*, 244-273.

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

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language, 55*(1), 37-58. https://doi.org/10.2307/412518

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica, 30*(3), 129-148. https://doi.org/10.1159/000259484

Hu 胡明扬, M. (1987). *北京话初探 Běijīng huà chūtàn* [On Beijing Mandarin]. Beijing: Commercial Press.

Hsu, Y.-Y., & German, J. S. (2018). Prosodic organization and focus realization in Taiwan Mandarin. In *Proceedings of the 6th International Symposium on Tonal Aspects of Languages (TAL 2018)* (pp. 24-28).

Hyman, L. M. (1977). On the nature of linguistic stress. *Studies in Stress and Accent (SCOPIL 4)*. Los Angeles: University of Southern California.

Hyman, L. M. (2016, May 24-27). Lexical vs. grammatical tone: Sorting out the differences. Paper presented at the *5th International Symposium on Tonal Aspects of Languages (TAL 2016)*, Buffalo, NY, United States.

Hyman, L., & Schuh, R. (1974). Universals of tone rules: Evidence from West Africa. *Linguistic Inquiry, 5*(1), 81-115.

Jenner, B. (1976). Interlanguage and foreign accent. *Interlanguage Studies Bulletin, 1*(2), 166-195.

Jesney, K. (2004). *The use of global accent rating in studies of L2 acquisition*. Calgary, AB: University of Calgary Language Research Center Reports.

Jiang, P., & Chen, A. (2011). Representation of Mandarin intonations: Boundary tone revisited. In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)* (Vol. 1, pp. 97-109).

Jilka, M. (2000). *The contribution of intonation to the perception of foreign accent* (Doctoral dissertation). University of Stuttgart.

Juffs, A. (1990). Tone, syllable structure, and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics in Language Teaching, 28*(2), 99-115. https://doi.org/10.1515/iral.1990.28.2.99

Juhász, K., & Bartos, H. (2023). Mandarin question intonation patterns in the production of Hungarian learners of Chinese. In *Proceedings of the International Congress of Phonetic Sciences (ICPHS 2023)* (pp. 1425-1429).

Jun, S.-A. (Ed.). (2005). *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press.

Jun, S.-A., & Fougeron, C. (2002). *Realizations of accentual phrase in French. Probus*, 14(1), 147-172.

Kabak, B., & Vogel, I. (2001). *The phonological word and stress in Turkish. Phonology*, 18(3), 315-360.

Kadota, S. (2007). *Shadowing to ondoku no kagaku* [The science of shadowing and oral reading]. Tokyo: Cosmopier.

Kadota, S. (2019). *Shadowing as a practice in second language acquisition: Connecting inputs and outputs*. London & New York: Routledge.

Kahane, J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy, 151*(1), 11-19. https://doi.org/10.1002/aja.1001510103

Kang, W., & Xu, Y. (2024). Tone-syllable synchrony in Mandarin: New evidence and implications. *Speech Communication, 163*, Article 103121. https://doi.org/10.1016/j.specom.2024.103121

Kirkpatrick, A., Deterding, D., & Wong, J. (2008). The international intelligibility of Hong Kong English. *World Englishes, 27*(3-4), 480-501. https://doi.org/10.1111/j.1467-971X.2008.00573.x

Kisilevsky, B. S., Hains, S. M. J., Lee, K., Xie, X., Huang, H., Ye, H. H., Zhang, K., & Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science, 14*(3), 220-224. https://doi.org/10.1111/1467-9280.02435

Krämer, M. (2009). *Syllable structure*. In *The phonology of Italian*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199290796.003.0005

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica, 55*(3-4), 243-276. https://doi.org/10.1556/ALing.55.2008.3-4.2

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684-686. https://doi.org/10.1126/science.277.5326.684

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511808814

Ladd, R., & Silverman, K. E. A. (1984). Vowel intrinsic pitch in connected speech. *Phonetica, 41*(1), 31-40. https://doi.org/10.1159/000261708

Ladefoged, P., & Johnson, K. (2010). *A course in phonetics* (6th ed.). Boston, MA: Wadsworth Cengage Learning.

Lane, H. (1963). Foreign accent and speech distortion. *Journal of the Acoustical Society of America, 35*(3), 451-453. https://doi.org/10.1121/1.1918501

Leben, W. R. (1973). *Suprasegmental phonology* (Doctoral dissertation). Massachusetts Institute of Technology.

Lecanuet, J. P. (1996). Prenatal auditory experience. In I. Deliege & J. Sloboda (Eds.), *Musical beginnings: Origins and development of musical competence* (pp. 3-34). Oxford: Oxford University Press.

Lee, J. H. N., Yip, K.-F., Liberman, M., & Kuang, J. (2024). Some prosodic consequences of varied discourse functions in a Cantonese sentence-final particle. In *Proceedings of Speech Prosody 2024* (pp. 632-636). https://doi.org/10.21437/SpeechProsody.2024-128

Lee, O. J. (2005). *The prosody of questions in Beijing Mandarin* (Doctoral dissertation). The Ohio State University.

Lee, Y.-C., Wang, T., & Liberman, M. (2016). Production and perception of Tone 3 focus in Mandarin Chinese. *Frontiers in Psychology, 7*, Article 1058. https://doi.org/10.3389/fpsyg.2016.01058

Lee-Kim, S.-I. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association*, *44*(3), 261–282. doi:10.1017/S0025100314000267

Levow, G.-A. (2004). Prosody-based topic segmentation for Mandarin broadcast news. In *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 137-140). https://aclanthology.org/N04-4035

Levow, G.-A. (2005). Turn-taking in Mandarin dialogue: Interactions of tone and intonation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing (IJCNLP 2005)*. https://aclanthology.org/I05-3010

Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech. In *Proceedings of Speech Prosody 2002* (pp. 39-46). https://doi.org/10.21437/SpeechProsody.2002-6

Li, A., & Wang, H. (2004). Friendly speech analysis and perception in Standard Chinese. In *Proceedings of Interspeech 2004* (pp. 897-900). https://doi.org/10.21437/Interspeech.2004-324

Li, A., Fang, Q., & Dang, J. (2011). Emotional intonation in a tone language: Experimental evidence from Chinese. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 17-21). Hong Kong.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.

Li Wen-Chao, (1999). *A Diachronically-Motivated Segmental Phonology of Mandarin Chinese*. New York, United States of America: Peter Lang Verlag.

Li, W. (2021). La querelle delle due forme del terzo tono: Tratti distintivi e pedagogie tonali a confronto. In C. Romagnoli & S. Conti (Eds.), *La lingua cinese in Italia: Studi su didattica e acquisizione* (pp. 205-222). Roma: Roma TrE-Press.

Li, W., & Yang, Y. (2009). Perception of prosodic hierarchical boundaries in Mandarin Chinese sentences. *Neuroscience, 158*(4), 1416-1425. https://doi.org/10.1016/j.neuroscience.2008.10.065

Li, X., Chen, Y., & Yang, Y. (2011). Immediate integration of different types of prosodic information during online spoken language comprehension: An ERP study. *Brain Research, 1386*, 139-152. https://doi.org/10.1016/j.brainres.2011.02.051

Li, Y. (2015, August 1). Tone sandhi and tonal coarticulation in Fuzhou Min. Paper presented at the *18th International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK.

Liberman, M. Y. (1975). *The intonational system of English* (Doctoral dissertation). Massachusetts Institute of Technology.

Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge: Cambridge University Press.

Lin 林茂灿, M. (2004). *Hànyǔ yǔdiào hé shēngdiào* 汉语语调和声调 [Chinese intonation and tone]. *Yǔyán Wénzì Yìngyòng 语言文字应用* [Applied Linguistics], 3, 57-67.

Lin 林茂灿, M. (2006). *Yíwèn hé chénshù yǔqì yǔ biānjiè diào* 疑问和陈述语气与边界调 [Interrogative vs. declarative and boundary tone]. *Zhōngguó Yǔwén 中国语文* [Studies of the Chinese Language], 4, 364-376.

Lindblom, B., Hardcastle, W., & Marchal, A. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (Vol. 55). Dordrecht: Springer. https://doi.org/10.1007/978-94-009-2037-8_16

Ling, B., & Liang, J. (2017). Focus encoding and prosodic structure in Shanghai Chinese. *The Journal of the Acoustical Society of America, 141*(6), EL610-EL616. https://doi.org/10.1121/1.4989739

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, M., Chen, Y., & Schiller, N. O. (2016). Online processing of tone and intonation in Mandarin: Evidence from ERPs. *Neuropsychologia, 91*, 307-317. https://doi.org/10.1016/j.neuropsychologia.2016.08.025

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods, 44*(4), 1042-1051. https://doi.org/10.3758/s13428-012-0203-3

Lu, Y., Aubergé, V., & Rilliard, A. (2012). Do you hear my attitude? Prosodic perception of social affects in Mandarin. In *Proceedings of Speech Prosody 2012 - 6th International Conference on Speech Prosody* (pp. 685-688). https://hal.science/hal-00744696

Mackey, W. F. (2000). The description of bilingualism. In L. Wei (Ed.), *The bilingualism reader* (pp. 26-54). Oxford: Routledge.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics, 26*(4), 381-400. https://doi.org/10.1006/jpho.1998.0081

Marotta, G., & Sardelli, E. (2009). *Prosodiatopia: Parametri prosodici per un modello di riconoscimento diatopico.* In *Atti del XL Congresso Internazionale di Studi della Società di Linguistica Italiana* (pp. 411–435).

Marotta, G., & Sorianello, P. (1999). *Question intonation in Sienese Italian.* In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 14)* (Vol. 2, pp. 1161–1164).

Marques, C., Moreno, S., Castro, S. L., & Besson, M. (2007). Musicians detect pitch violations in speech better than non-musicians: Behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience, 19*(9), 1453-1463.

McCarthy, J. J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology 1993* (pp. 79-153). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-3712-8_4

McGory, J. T. (1997). *Acquisition of intonational prominence in English by Seoul Korean and Mandarin Chinese speakers* (Doctoral dissertation). The Ohio State University.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics, 32*(4), 543-563. https://doi.org/10.1016/j.wocn.2004.02.002

Mennen, I. (2007). Phonological and phonetic influences in non-native intonation. In J. Trouvain & U. Gut (Eds.), *Non-native prosody: Phonetic description and teaching practice* (pp. 53-76). Berlin: Mouton de Gruyter.

Mennen, I. (2015). Beyond segments: Towards an L2 intonation learning theory. In E. Delais-Roussarie, M. Avanzi, & S. Herment (Eds.), *Prosody, phonology and phonetics: Prosody and language in contact* (pp. 171-188). Dordrecht: Springer.

Mennen, I., Chen, A., & Karlsson, F. (2010). Characterising the internal structure of learner intonation and its development over time. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech* (pp. 319-324). Poznań: Adam Mickiewicz University.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97. https://doi.org/10.1037/h0043158

Morbiato, A. (2020). *Il tema in cinese tra frase e testo: Struttura sintattica, informativa e del discorso* (1st ed.). Venezia: Cafoscarina.

Moschetti, A. (2007). Dalla voce materna al cervello del neonato. *Quaderni ACP, 14*(4), 188-189.

Mueller-Liu, P. (2006). Signalling affect in Mandarin Chinese – The role of non-lexical utterance-final edge tones. *Speech Prosody 2006*, Paper 048-0. https://doi.org/10.21437/SpeechProsody.2006-146

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520-531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*(3), 316-327. https://doi.org/10.1017/S0261444811000103

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). Measuring the facets of musicality: The Goldsmiths Musical Sophistication Index (Gold-MSI). *Personality and Individual Differences, 60*(Supplement), S35. https://doi.org/10.1016/j.paid.2013.07.041

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

Nguyen, T. A. T., Ingram, J. C. L., & Pensalfini, J. R. (2008). Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. *Journal of Phonetics, 36*(1), 158-190. https://doi.org/10.1016/j.wocn.2007.09.001

Nolan, F. (2006). Intonation. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 433-457). Oxford: Blackwell.

Nye, P., & Fowler, C. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics, 31*(1), 63-79. https://doi.org/10.1016/S0095-4470(02)00073-5

O'Brien, M., & Gut, U. (2010). Phonological and phonetic realisation of different types of focus in L2 speech. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *Achievements*

and perspectives in the acquisition of second language speech: New Sounds 2010 (pp. 205-215). Frankfurt: Peter Lang.

Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica, 40*(1), 1-18. https://doi.org/10.1159/000261678

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F_0 of voice. *Phonetica, 41*(1), 1-16. https://doi.org/10.1159/000261706

Ortega, L. (2009). *Understanding second language acquisition* (1st ed.). London: Routledge. https://doi.org/10.4324/9780203777282

Ouyang, I. C., & Kaiser, E. (2013). Prosody and information structure in a tone language: An investigation of Mandarin Chinese. *Language, Cognition and Neuroscience, 30*(1-2), 57-72. https://doi.org/10.1080/01690965.2013.805795

Pan, H.-H. (2007). Focus and Taiwanese unchecked tones. In C. Lee, M. K. Gordon, & D. Büring (Eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation* (pp. 195-213). Dordrecht: Springer.

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology, 2*, 142.

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion, 28*(2), 230-244. https://doi.org/10.1080/02699931.2013.812033

Pell, M. D. (2006). Judging emotion and attitudes from prosody following brain damage. *Progress in Brain Research, 156*, 303-317. https://doi.org/10.1016/S0079-6123(06)56017-0

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37*(4), 417-435. https://doi.org/10.1016/j.wocn.2009.07.005

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition, 41*(1), 59-86. https://doi.org/10.1017/S0272263117000444

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2021). Even in the best-case scenario L2 learners have persistent difficulty perceiving and utilizing tones in Mandarin: Findings from

behavioral and event-related potentials experiments. *Studies in Second Language Acquisition, 43*(2), 268-296. https://doi.org/10.1017/S027226312000039X

Peng, S. (1997). Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics, 25*(3), 371-400. https://doi.org/10.1006/jpho.1997.0047

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* (1st ed., pp. 230-270). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.003.0009

Piccardo, E. (2016). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Phonological scale revision process report.* Strasbourg: Council of Europe.

Piccardo, E., & North, B. (2017). Developing phonology descriptors for the Common European Framework of Reference (CEFR). In M. O'Brien & J. Levis (Eds.), *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, Calgary, AB, August 2016 (pp. 97-109). Ames, IA: Iowa State University.

Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics, 26*, 219-233. https://doi.org/10.1017/S0267190506000110

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation). Massachusetts Institute of Technology.

Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (pp. 271-311). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3839.003.0016

Pike, K. L. (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

Pike, K. L. (1948). *Tone languages: A technique for determining the number and pitch contrasts in a language, with studies in tonemic substitution and fusion.* Ann Arbor: University of Michigan Press.

Prom-on, S., Liu, F., & Xu, Y. (2012). Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *The Journal of the Acoustical Society of America, 132*(1), 421-432. https://doi.org/10.1121/1.4725762

Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., & Crèpin, G. (1988). Fetal hearing. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 28*(3), 191-212. https://doi.org/10.1016/0028-2243(88)90007-9

Raini, E., & Wang, R. (2023). *La pronuncia del cinese: Teoria ed esercizi.* Milano: Hoepli.

Refinetti, R. (2016). *Circadian physiology* (3rd ed.). Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781315370268

Ren, G.-Q., Tang, Y.-Y., Li, X.-Q., & Sui, X. (2013). Pre-attentive processing of Mandarin tone and intonation: Evidence from event-related potentials. In F. Signorelli & D. Chirchiglia (Eds.), *Functional brain mapping and the endeavor to understand the working brain* (pp. 125-140). Rijeka: InTech. https://doi.org/10.5772/56503

Ren, G.-Q., Yang, Y., & Li, X. (2009). Early cortical processing of linguistic pitch patterns as revealed by the mismatch negativity. *Neuroscience, 162*(1), 87-95. https://doi.org/10.1016/j.neuroscience.2009.04.021

Rialland, A. (2007). Question prosody: An African perspective. In T. Riad & C. Gussenhoven (Eds.), *Tones and tunes: Typological studies in word and sentence prosody* (Vol. 1, pp. 35-62). Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110207569.35

Roach, P. J. (2001). *Phonetics.* Oxford: Oxford University Press.

Sacks, O. (2008). *Musicophilia: Tales of music and the brain.* New York: Vintage Books.

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of native and non-native phonetic contrasts by musicians and tone language speakers. *Acta Psychologica, 138*(1), 1-10.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech Communication, 46*(3-4), 405-417. https://doi.org/10.1016/j.specom.2005.01.010

Santangelo, M., Persici, V., Caricati, L., Corsano, P., Gordon, R. L., & Majorano, M. (2023). The adaptation and validation of the Goldsmiths Musical Sophistication Index (Gold-MSI) in Italian: The Gold-MSI-IT. *Psychology of Music.* Advance online publication. https://doi.org/10.1177/03057356231204855

Santiago-Vargas, F., & Delais-Roussarie, E. (2012). La prosodie des énoncés interrogatifs en français L2. In L. Besacier, B. Lecouteux, & G. Sérasset (Eds.), *Actes des Journées d'études sur la Parole JEP/TALN 2012* (pp. 265-272). Grenoble: AFCP/ATALA.

Sapir, S. (1989). The intrinsic pitch of vowels: Theoretical, physiological, and clinical considerations. *Journal of Voice, 3*(1), 44-51. https://doi.org/10.1016/S0892-1997(89)80121-3

Sardelli, E., & Marotta, G. (2007). *Prosodic parameters for the detection of regional varieties in Italian.* In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 16)* (pp. 1281–1284).

Saville-Troike, M. (2012). *Introducing second language acquisition* (2nd ed.). Cambridge: Cambridge University Press.

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

Sbranna, S., Ventura, C., Albert, A., & Grice, M. (2023). Prosodic marking of information status in Italian. *Journal of Phonetics, 97,* 101371. https://doi.org/10.1016/j.wocn.2023.101371

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92. https://doi.org/10.1177/0022022101032001009

Scholz, F., & Chen, Y. (2014). The independent effects of prosodic structure and information status on tonal coarticulation: Evidence from Wenzhou Chinese. In J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller, & E. Van Zanten (Eds.), *Above and beyond the segments* (pp. 275-287). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.189.22sch

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Selkirk, E. (2003). Sentence phonology. In *International encyclopedia of linguistics* (Vol. 3, pp. 150-153). Oxford: Oxford University Press.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (Vol. 2, pp. 867-870).

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193-247. https://doi.org/10.1007/BF01708572

Shen, C., & Xu, Y. (2016). Prosodic focus with post-focus compression in Lan-Yin Mandarin. *Speech Prosody 2016,* 340-344. https://doi.org/10.21437/SpeechProsody.2016-70

Shen, J. (1992). On Chinese intonation models. *Chinese Studies, 4,* 16-24.

Shen, X. S. (1990a). *The prosody of Mandarin Chinese.* Berkeley: University of California Press.

Shen, X. S. (1990b). Tonal coarticulation in Mandarin. *Journal of Phonetics, 18*(2), 281-295. https://doi.org/10.1016/S0095-4470(19)30394-8

Shih, C. (1997). Mandarin third tone sandhi and prosodic structure. In J. Wang & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 81-124). Berlin: De Gruyter. https://doi.org/10.1515/9783110822014.81

Shih, C.-L. (1988). Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory, 3,* 83-109.

Silverman, K. (1986). F_0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica, 43*(1-3), 76-91. https://doi.org/10.1159/000261762

Sparvoli, C. (2017). From phonological studies to teaching Mandarin tone. In I. Kecskes & C. Sun (Eds.), *Key issues in Chinese as a second language research* (pp. 81-100). New York: Routledge. https://doi.org/10.4324/9781315660264-4

Speer, S. R., Shih, C.-L., & Slowiaczek, M. L. (1989). Prosodic structure in language understanding: Evidence from tone sandhi in Mandarin. *Language and Speech, 32*(4), 337-354. https://doi.org/10.1177/002383098903200403

Steele, S. A. (1986). Interaction of vowel F_0 and prosody. *Phonetica, 43*(1-3), 92-105. https://doi.org/10.1159/000261763

Stewart, J. M. (1965). The typology of the Twi tone system. *Bulletin of the Institute of African Studies, 1,* 1-27.

Stockwell, R. P., & Bowen, D. (1965). *The sounds of English and Spanish.* Chicago, IL: University of Chicago Press.

Sun, S. H. (1998). *The development of a lexical tone phonology in American adult learners of Standard Mandarin Chinese.* Honolulu: University of Hawaii Press.

Swerts, M., & Zerbian, S. (2010). Intonational differences between L1 and L2 English in South Africa. *Phonetica, 67*(3), 127-146. https://doi.org/10.1159/000319366

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30. https://doi.org/10.1017/S0272263106060013

Třísková, H. (2011). The structure of the Mandarin syllable: Why, when and how to teach it. *Archiv Orientální, 79*(1), 99-134.

Třísková, H. (2019). Is the glass half-full, or half-empty? The alternative concept of stress in Mandarin Chinese. *Studies in Prosodic Grammar, 4*(2), 64-105.

Tseng, C.-Y. (2002). The prosodic status of breaks in running speech: Examination and evaluation. In *Proceedings of Speech Prosody 2002* (pp. 667-670). Aix-en-Provence, France.

Tseng, C.-Y., Pin, S., Lee, Y., Wang, H., & Chen, Y. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication, 46*(3-4), 284-309. https://doi.org/10.1016/j.specom.2005.03.015

Tu, J.-Y., Hsiung, Y., Cha, J.-H., Wu, M.-D., & Sung, Y.-T. (2016, May 31-June 3). Tone production of Mandarin disyllabic words by Korean learners. In *Proceedings of Speech Prosody 2016* (pp. 375-379). Boston, MA, USA.

Ueyama, M., & Jun, S.-A. (1998). Focus realization in Japanese English and Korean English intonation. In *Japanese-Korean Linguistics* (Vol. 7, pp. 629-645). Stanford: CSLI Publications.

Ulbrich, C. (2008). Acquisition of regional pitch patterns in L2. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Speech Prosody 2008 Conference* (pp. 575-578). Campinas: State University of Campinas.

Vance, T. J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetica, 33*(5), 368-392. https://doi.org/10.1159/000259793

Vetchinnikova, S., Huhtala, A., & Yangarber, R. (2023). Chunking up speech in real time: Linguistic predictors and cognitive constraints. *Language and Cognition, 15*(2), 299-334. https://doi.org/10.1017/langcog.2023.15

Vigliano, D., Yoshimoto, K., & Pellegrino, E. (2016). A self-imitation technique for the improvement of prosody in L2 Italian. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing* (pp. 1189-1192). Tohoku.

Vogel, I. (2009). *The status of the clitic group.* In J. Grijzenhout & B. Kabak (Eds.), *Phonological domains: Universals and deviations* (pp. 15-46). Berlin: Mouton de Gruyter.

Wang, A. (2003). *Research on the pitch downtrend of intonation in Putonghua.* Beijing University.

Wang, B., & Xu, Y. (2011). Differential prosodic encoding of topic and focus in sentence-initial position in Mandarin Chinese. *Journal of Phonetics, 39*(4), 595-611. https://doi.org/10.1016/j.wocn.2011.03.006

Wang, B., Wang, L., & Qadir, T. (2011). Prosodic realization of focus in six languages/dialects in China. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII),* 144-147. Hong Kong.

Wang, J. (2003). *Rhythmic grouping, tone sandhi and stress in Beijing Mandarin.* Macquarie University.

Wang, L. 王力 (1958). *Wáng Lì. Hànyǔ shīlǜ xué 汉语诗律学 [Chinese versification].* Shanghai: Xin Zhishi Chubanshe.

Wang, L. 王力 (1980). *Hànyǔ shǐgǎo* (漢語史稿) *[History of the Chinese language].* Beijing: Commercial Press. ISBN 978-7-101-01553-9.

Wang, P. 王萍, & Shi, F. 石锋. (2011). *Hànyǔ yǔdiào de jīběn móshì 汉语语调的基本模式 [Intonation patterns in Mandarin Chinese]. Nánkāi dàxué xuébào 南开大学学报,* 2, 1-11.

Wang, X. (2013). Perception of Mandarin tones: The effect of L1 background and training. *Journal of Second Language Pronunciation, 1*(2), 85-108.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America, 106*(6), 3649-3658. https://doi.org/10.1121/1.428217

Wee, L.-H. (2022). Tonal processes conditioned by morphosyntax. In C.-R. Huang, I.-H. Chen, Y.-H. Lin, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics* (pp. 313-335). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108329019.018

Welmers, W. E. (1959). Tonemics, morphotonemics, and tonal morphemes. *General Linguistics, 4,* 1-9.

Welmers, W. E. (1974). *African language structures.* Berkeley: University of California Press.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition, 20*(1), 1-25. https://doi.org/10.1017/S0272263100001343

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F₀ of vowels. *Journal of Phonetics, 23*(3), 349-366. https://doi.org/10.1016/S0095-4470(95)80165-0

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49,* 25-47. https://doi.org/10.1159/000261911

Wheeldon, L. (2013). Generating prosodic structure. In *Aspects of language production* (pp. 249-274). London: Psychology Press. https://doi.org/10.4324/9781315804453

White, C. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of the Chinese Language Teachers Association, 16*(2), 27-56.

Wichmann, A. (2000). The attitudinal effects of prosody and how they relate to emotion. In *Proceedings of Speech Emotion 2000* (pp. 143-148). Newcastle, Northern Ireland.

Willems, N. (1982). *English intonation from a Dutch point of view.* Dordrecht: Foris.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*(4), 420-422.

Wu, Z. 吴宗济. (1982). *Wú Zōngjì. Pǔtōnghuà yǔjù zhōng de shēngdiào biànhuà 普通话语句中的声调变化 [Tonal changes in Mandarin sentences]. Zhōngguó yǔwén 中国语文,* 171, 439-449.

Wu, W., & Xu, Y. (2010). Prosodic focus in Hong Kong Cantonese without post-focus compression. *Speech Prosody 2010,* Paper 040. https://doi.org/10.21437/SpeechProsody.2010-85

Xin, Y., & Zhang, W. (2009). *Mǔyǔ zhòngyīn duì Yìdàlì liúxuéshēng Hànyǔ shēngdiào xídé de yǐngxiǎng 母语重音对意大利留学生汉语声调习得的影响 [The influence of native*

language stress on the acquisition of Chinese tones by Italian students]. Wénhuà jiēchù yǔ Běijīnghuà biānyì yánjiū 文化接触与北京话变异研究, 121-137.

Xu, C. X., & Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association, 33*(2), 165-181. https://doi.org/10.1017/S0025100303001270

Xu, H. (2019). The effectiveness of visualizing pitch curves for Chinese. In I. Sagiyama & M. Castorina (Eds.), *Trajectories: Selected papers in East Asian studies* (*Florentalia*). Firenze University Press.

Xu, L. (2006). Topicalization in Asian languages. In M. Everaert & H. C. Riemsdijk (Eds.), *The Wiley Blackwell companion to syntax* (2nd ed., pp. 1-30). Wiley. https://doi.org/10.1002/9781118358733.wbsyncom024

Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America, 95*(4), 2240-2253. https://doi.org/10.1121/1.408684

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61-83. https://doi.org/10.1006/jpho.1996.0034

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F_0 contours. *Journal of Phonetics, 27,* 55-105.

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. (2013). ProsodyPro – A tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7-10). Aix-en-Provence, France.

Xu, Y., & Lee, A. (2022). Tonal processes defined as articulatory-based contextual tonal variation. In C.-R. Huang, I.-H. Chen, Y.-H. Lin, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics* (pp. 275-290). Cambridge University Press. https://doi.org/10.1017/9781108329019.016

Xu, Y., & Prom-on, S. (2019). Economy of effort or maximum rate of information? Exploring basic principles of articulatory dynamics. *Frontiers in Psychology, 10,* 2469. https://doi.org/10.3389/fpsyg.2019.02469

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America, 111*(3), 1399-1413. https://doi.org/10.1121/1.1445789

Xu, Y., & Wang, E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33,* 319-337. https://doi.org/10.1016/S0167-6393(00)00063-7

Xu, Y., & Wang, M. (2009). Organizing syllables into groups – Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics, 37*(4), 502-520. https://doi.org/10.1016/j.wocn.2009.08.003

Xu, Y., & Wang, Q. E. (1997). What can tone studies tell us about intonation? In *Proceedings of the ESCA Workshop on Intonation* (pp. 337-340). Athens, Greece.

Xu, Y., Chen, S., & Wang, B. (2012). Prosodic focus with and without post-focus compression: A typological divide within the same language family? *The Linguistic Review, 29*(1), 131-147. https://doi.org/10.1515/tlr-2012-0006

Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. In S. Hancil & D. Hirst (Eds.), *Iconicity in language and literature* (Vol. 13, pp. 33-60). John Benjamins. https://doi.org/10.1075/ill.13.02xu

Yang, B., & Yang, N. (2017). Development of disyllabic tones in different learning contexts. *International Review of Applied Linguistics in Language Teaching, 57*(2), 237-266. https://doi.org/10.1515/iral-2016-0004

Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice.* John Benjamins Publishing Company.

Yang, C. (2019). Teaching Chinese intonation and rhythm. In C. Shei, M. E. McLellan Zikpi, & D.-L. Chao (Eds.), *The Routledge handbook of Chinese language teaching* (pp. 180-194). Abingdon & New York: Routledge.

Yang, L. (1995). Yáng Lì 杨莉. *Sān zhǒng yíwèn jù de yǔdiào zhī yítóng 三种疑问句的语调之异同 [An intonational comparison of three types of interrogatives]. Zhōngguó yǔxué 中国语学,* 144-152.

Yang, X., & Yang, Y. (2012). Prosodic realization of rhetorical structure in Chinese discourse. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(4), 1196-1206. https://doi.org/10.1109/TASL.2011.2173676

Yang, Y., & Wang, B. (2002). Acoustic correlates of hierarchical prosodic boundary in Mandarin. In *Proceedings of Speech Prosody 2002* (pp. 707-710). https://doi.org/10.21437/SpeechProsody.2002-162

Yi, X. (2004). Separation of functional components of tone and intonation from observed F0 patterns. In *Traditional phonology to modern speech processing: Festschrift for Professor Wu Zongji's 95th birthday* (pp. 485-505). Beijing: Foreign Language Teaching and Research Press.

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

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

Yuan, J., Shih, C., & Kochanski, G. P. (2002). Comparison of declarative and interrogative intonation in Chinese. In *Proceedings of Speech Prosody 2002* (pp. 711-714). https://doi.org/10.21437/SpeechProsody.2002-163

Zhang, H. (2007). *A phonological study of second language acquisition of Mandarin Chinese tones* (Doctoral dissertation). The University of North Carolina at Chapel Hill.

Zhang, H. (2010). Phonological universals and tone acquisition. *Journal of the Chinese Language Teachers Association, 45*(1), 39-65.

Zhang, H., & Qian, Y. (Eds.). (2020). *Prosodic studies: Challenges and prospects* (1st ed.). Routledge.

Zhang, J. (2014). Tones, tonal phonology, and tone sandhi. In C.-T. J. Huang, Y.-H. A. Li, & A. Simpson (Eds.), *The handbook of Chinese linguistics* (pp. 443-464). Wiley. https://doi.org/10.1002/9781118584552.ch17

Zhang, J. (2022). Tonal processes defined as tone sandhi. In C.-R. Huang, I.-H. Chen, Y.-H. Lin, & Y.-Y. Hsu (Eds.), *The Cambridge handbook of Chinese linguistics* (pp. 291-312). Cambridge University Press. https://doi.org/10.1017/9781108329019.017

Zheng, T., Levelt, C. C., & Chen, Y. (2025). The affective iconicity of lexical tone: Evidence from Standard Chinese. *The Journal of the Acoustical Society of America, 157*(1), 396-408. https://doi.org/10.1121/10.0034863

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*(1), 69-84. https://doi.org/10.1016/j.system.2007.11.004

Zou, T., Caspers, J., & Chen, Y. (2022). Perception of different tone contrasts at sub-lexical and lexical levels by Dutch learners of Mandarin Chinese. *Frontiers in Psychology, 13,* 283139. https://doi.org/10.3389/fpsyg.2022.828313

Appendix A: Informed consent

Benvenuto in questo esperimento sull'acquisizione della lingua cinese da parte di apprendenti italofoni! Sono Davide Francolino, responsabile scientifico del progetto. In questo documento troverai due sezioni:
1) **Informazioni sullo svolgimento delle attività sperimentali**;
2) **Informativa sulla tutela della privacy (da restituire firmata)**.

INFORMAZIONI SULLO SVOLGIMENTO DELLE ATTIVITÀ SPERIMENTALI

La partecipazione alla ricerca prevede due sezioni distinte:

1) **Attività in presenza** (circa **40 minuti**)

L'attività in presenza prevede la lettura a coppie di brevi dialoghi in cinese con caratteri semplificati e pinyin. L'attività è suddivisa in quattro blocchi di ca. 30 dialoghi, intervallati da brevi pause di ca. 3 minuti ciascuna. Prima di leggere, i partecipanti possono prendersi del tempo per familiarizzare con il testo. Se necessario, il partecipante può ripetere la lettura a propria discrezione. L'attività sarà registrata in formato audio.

Si consiglia di portare dell'acqua per restare idratati durante la lettura.

2) **Attività online** (circa **20 minuti**)

L'attività online comprende un breve test di competenza linguistica (ca. 10 min.) e un questionario su Google Moduli (ca. 10 min.)

Il test di competenza linguistica si svolgerà da remoto, alla presenza del ricercatore. Il questionario è invece da svolgersi online in maniera autonoma. Sia il test di competenza linguistica che il questionario su Google Moduli saranno anonimizzati, non richiederanno alcuna preparazione preliminare e avranno esclusivamente uno scopo scientifico, volto a delineare il profilo sociolinguistico dei partecipanti.

Compenso

Ogni partecipante riceverà un compenso di 8€ (via Paypal, Revolut o bonifico bancario). Il pagamento non sarà immediato, ma verrà approvato previo verifica della qualità dei dati inviati dal partecipante, entro un massimo di 3 giorni dal termine di tutte le attività.

Appendix Figure 1 Informed consent, p.1

INFORMATIVA SULLA TUTELA DELLA PRIVACY

Tutti i dati raccolti verranno elaborati e divulgati in forma aggregata, secondo la legge sulla privacy. La pianificazione della gestione dei dati raccolti verrà effettuata in ottemperanza della normativa relativa alla protezione dei dati personali (Regolamento UE 2016/679 - Regolamento Generale sulla Protezione dei Dati - GDPR - applicativo dal 25 maggio 2018 e in conformità al Decreto Legislativo 30 giugno 2003 n. 196 Codice in materia di protezione dei dati personali). I risultati della ricerca potranno essere oggetto di pubblicazione, di condivisione con comitati di riviste scientifiche italiane e internazionali al fine di controllare che la ricerca sia condotta correttamente e in conformità alle disposizioni vigenti, nonché con uditori in convegni in Italia e all'estero; i dati saranno sempre divulgati in forma aggregata e anonima, in modo che il Partecipante non sia identificabile.
La durata complessiva stimata per la partecipazione al presente studio è di 60 minuti. Tuttavia, non sono previsti vincoli di tempo obbligatori, pertanto il Partecipante potrà svolgere le attività secondo il ritmo che ritiene più confacente alle proprie esigenze. Lo studio non comporta alcun rischio né conseguenza negativa per la salute psicologica o fisica del Partecipante. Qualora interessati, sarà possibile contattare tramite e-mail il Ricercatore responsabile per richiedere, quando disponibili, i risultati della ricerca.

Responsabile scientifico: Davide Francolino, Università per Stranieri di Siena

Contatto e-mail: davide.francolino@unistrasi.it

CATEGORIE DI DATI PERSONALI, FINALITÀ E BASE GIURIDICA

Il trattamento ha ad oggetto l'acquisizione di alcuni dati personali e di registrazioni audio da parte del Partecipante. I dati raccolti saranno analizzati al fine di eseguire le attività di ricerca scientifica previste dal progetto di ricerca di dottorato del Ricercatore responsabile. Il trattamento dei dati personali verrà effettuato con strumenti informatici, adottando misure tecniche e organizzative adeguate a proteggerli da accessi non autorizzati o illeciti, dalla distruzione, e dalla perdita d'integrità e riservatezza, anche accidentali.

OBIETTIVO DELLO STUDIO

La presente ricerca si propone di esaminare alcune problematiche relative all'acquisizione della lingua cinese da parte di apprendenti di madrelingua italiana.

TEMPI E MODALITÀ DI CONSERVAZIONE DEI DATI

I dati raccolti saranno conservati per tutta la durata della ricerca di tesi e successivamente archiviati per un periodo di 5 anni. Al termine di questo periodo, le risposte ai questionari e le registrazioni audio, debitamente anonimizzate, potranno essere mantenute per ulteriori progetti di ricerca futuri. La conservazione dei dati sarà effettuata tramite la piattaforma Google Drive istituzionale del Ricercatore.

RITIRO DALLO STUDIO E/O REVOCA DELL'AUTORIZZAZIONE

È possibile rifiutarsi di partecipare allo studio o ritirare il proprio consenso in qualsiasi momento, senza alcuna conseguenza (vedi "diritto all'oblio"). Qualora il Partecipante desideri revocare la propria autorizzazione è pregato di informare il Ricercatore senza alcuna formalità tramite contatto e-mail.

Appendix Figure 2 Informed consent, p.2

LIBERATORIA PER RIPRESE AUDIO

Ad integrazione dell'informativa di cui sopra, La informiamo che, previa acquisizione del suo consenso ai sensi dell'art. 96 e ss. della Legge n. 633/1941, le tracce audio raccolte nell'ambito della ricerca di cui sopra verranno utilizzate per finalità di ricerca scientifica. Il Partecipante autorizza il Ricercatore a riprodurre e condividere le tracce audio esclusivamente in forma aggregata e anonima, e solo quando strettamente necessario in ambiti accademici, ad esempio durante presentazioni a conferenze o per la valutazione da parte di comitati scientifici in vista di pubblicazioni. Le tracce audio potrebbero inoltre essere pubblicate su piattaforme online di Open Science, anche nell'ottica della creazione di un corpus di parlato Open Source. Resta inteso che, con la presente autorizzazione, il Partecipante vieta l'uso delle registrazioni contententi la propria voce in contesti che possano pregiudicare la dignità personale e il decoro.

Accordo per la Partecipazione
Firmando qui sotto, acconsento volontariamente a partecipare a questo studio:

_____ _____

(luogo e data) (firma del Partecipante)

Firma del Ricercatore responsabile: _____

Appendix Figure 3 Informed consent, p.3

Appendix B: Stimuli

1. Stim_T1T1_foc1

A: 应该说“喝粥”还是“吃粥”？

A: yīnggāi shuō “hē zhōu” háishì “chī zhōu”?

B: 喝粥。

B: hē zhōu.

A: 喝粥？

A: hē zhōu?

B: 对啊，粥是稀的嘛。

B: duì a, zhōu shì xī de ma.

2. Stim_T1T1_foc2

A: 你要喝粥还是喝酒？

A: nǐ yào hē zhōu háishì hē jiǔ?

B: 喝粥。

B: hē zhōu.

A: 喝粥？

A: hē zhōu?

B: 今天我肚子不太舒服。

B: jīntiān wǒ dùzi bù tài shūfu.

3. Stim_T1T2_foc1

A: “喝茶”还是“饮茶”，哪个说法更常用？

A: "hē chá" háishì "yǐn chá", nǎge shuōfǎ gèng chángyòng?

B: 喝茶。

B: hē chá.

A: 喝茶？

A: hē chá?

B: 对，“饮茶”这个说法比较仪式化。

B: duì, "yǐn chá" zhège shuōfǎ bǐjiào yíshìhuà.

4. Stim_T1T2_foc2

A: 在意大利，饭后你一般喝茶还是喝咖啡？

A: zài Yìdàlì, fàn hòu nǐ yìbān hē chá háishì hē kāfēi?

B: 喝茶。

B: hē chá.

A: 喝茶？

A: hē chá?

B: 对，不过在意大利人们喝咖啡比较常见。

B: duì, bù guò zài Yìdàlì rénmen hē kāfēi bǐjiào chángjiàn.

5. Stim_T1T3_foc1

A: 我多问一句，你今天喝不喝酒？

A: wǒ duō wèn yī jù, nǐ jīntiān hē bù hē jiǔ?

B: 喝酒。

B: hē jiǔ.

A: 喝酒？

A: hē jiǔ?

B: 你放心，我病已经好了。

B: nǐ fàngxīn, wǒ bìng yǐjīng hǎo le.

6. Stim_T1T3_foc2

A 你吃饭的时候一般喝酒还是喝水？

A: nǐ chīfàn de shíhou yìbān hē jiǔ háishì hē shuǐ?

B: 喝酒。

B: hē jiǔ.

A: 喝酒？

A: hē jiǔ?

B: 我知道这种习惯不太健康。

B: wǒ zhīdào zhè zhǒng xíguàn bù tài jiànkāng.

7. Stim_T1T4_foc1

A: 你的面里加不加菜？

A: nǐ de miàn lǐ jiā bu jiā cài?

B: 加菜。

B: jiā cài.

A: 加菜？

A: jiā cài?

B: 对，以前不加，今天可以。

B: duì, yǐqián bù jiā, jīntiān kěyǐ.

8. Stim_T1T4_foc2

A:你要加菜还是加面？

A: nǐ yào jiā cài háishì jiā miàn?

B: 加菜。

B: jiā cài.

A: 加菜？

A: jiā cài?

B: 对，以前加面，可是医生说我要少吃点面。

B: duì, yǐqián jiā miàn, kěshì yīshēng shuō wǒ yào shǎo chī diǎn miàn.

9. Stim_T2T1_foc1

A: 应该说“学车”还是“习车”？

A: yīnggāi shuō “xuéchē” háishì “xíchē”?

B: 学车。

B: xué chē.

A: 学车？

A: xué chē?

B: 对啊，没听过 “习车”，你从哪儿听来的？

B: duì a, méi tīng guò "xí chē", nǐ cóng nǎr tīng lái de?

10. Stim_T2T1_foc2

A: 暑假打算学车还是学英语？

A: shǔjià dǎsuàn xué chē háishì xué yīngyǔ?

B: 学车。

B: xué chē.

A: 学车？

A: xué chē?

B: 对，我本来要补习英语，但我决定等到暑假后再学。

B: duì, wǒ běnlái yào bǔxí Yīngyǔ, dàn wǒ juédìng děngdào shǔjià hòu zài xué.

11. Stim_T2T2_foc1

A: “读博”还是“念博”，普通话怎么说？

A: "dú bó" háishì "niàn bó", pǔtōnghuà zěnme shuō?

B: 读博。

B: dú bó.

A: 读博？

A: dú bó?

B: 对，没听过 “念博” ，你从哪儿听来的？

B: duì, méi tīng guo "niàn bó", nǐ cóng nǎr tīng lái de?

12. Stim_T2T2_foc2

A: 他在读博还是读研？

A: tā zài dú bó háishì dú yán?

B: 读博。

B: dú bó.

A: 读博？

A: dú bó？

B: 对啊，他研究生早就毕业了。

B: duì a, tā yánjiūshēng zǎo jiù bìyè le.

13. Stim_T2T3_foc1

A: 她负责提水还是倒水？

A: tā fùzé tí shuǐ háishì dào shuǐ?

B: 提水。

B: tí shuǐ.

A: 提水？

A: tí shuǐ?

B: 对，这个工作很辛苦。

B: duì, zhège gōngzuò hěn xīnkǔ.

14. Stim_T2T3_foc2

A: 你要用这个铁桶提水还是提沙子？

A: nǐ yào yòng zhè gè tiětǒng tí shuǐ háishì tí shāzi？

B: 提水。

B: tí shuǐ.

A: 提水？

A: tí shuǐ?

B: 对，我这边没有别的水桶，只能用它了。

B: duì, wǒ zhè biān méiyǒu bié de shuǐtǒng, zhǐ néng yòng tā le.

15. Stim_T2T4_foc1

A: 你择(zhái)菜还是洗菜？

A: nǐ zhái cài háishì xǐ cài?

B: 择菜。

B: zhái cài.

A: 择菜？

A: zhái cài?

B: 怎么，你以为我不会择菜吗？

B: zěnme, nǐ yǐwéi wǒ bù huì zhái cài ma?

16. Stim_T2T4_foc2

A: 应该说“择(zhái)菜”还是“择叶”？

A: yīnggāi shuō "zhái cài" háishì "zhái yè"?

B: 择菜。

B: zhái cài.

A: 择菜？

A: zhái cài?

B: 对啊，没听过 “择叶” ，你从哪儿听来的？

B: duì a, méi tīngguò "zhái yè", nǐ cóng nǎr tīng lái de?

17. Stim_T3T1_foc1

A: 一般说“煮粥”还是“煲粥”？

A: yībān shuō "zhǔ zhōu" háishì "bāo zhōu"?

B: 煮粥。

B: zhǔ zhōu.

A: 煮粥？

A: zhǔ zhōu?

B: 对啊，好像只有广东人才会说 “煲粥” 。

B: duì a, hǎoxiàng zhǐ yǒu Guǎngdōng rén cái huì shuō "bāo zhōu".

18. Stim_T3T1_foc2

A: 你是在煮粥还是煮米饭？

A: nǐ shì zài zhǔ zhōu háishì zhǔ mǐfàn?

B: 煮粥。

B: zhǔ zhōu.

A: 煮粥？

A: zhǔ zhōu?

B: 对，我肠胃不好，吃点流食。

B: duì, wǒ chángwèi bù hǎo, chī diǎn liúshí.

19. Stim_T3T2_foc1

A: 你喜欢煮茶还是泡茶？

A: nǐ xǐhuān zhǔ chá háishì pào chá?

B: 煮茶。

B: zhǔ chá.

A: 煮茶？

A: zhǔ chá?

B: 对，虽然一般人喜欢泡茶，但我觉得煮的茶更浓郁。

B: duì, suīrán yìbān rén xǐhuān pào chá, dàn wǒ juédé zhǔ de chá gèng nóngyù.

20. Stim_T3T2_foc2

A: 这个锅你用来煮汤还是煮茶？

A: zhège guō nǐ yòng lái zhǔ tāng háishì zhǔ chá?

B: 煮茶。

B: zhǔ chá.

A: 煮茶？

A: zhǔ chá?

B: 对啊，煮汤的话它太小了呀。

B: duì a, zhǔ tāng de huà tā tài xiǎo le ya.

21. Stim_T3T3_foc1

A: 你给猫洗脚还是擦脚？

A: nǐ gěi māo xǐ jiǎo háishì cā jiǎo?

B: 洗脚。

B: xǐ jiǎo.

A: 洗脚？

A: xǐ jiǎo?

B: 对，不过我每周只给它洗一次。

B: duì, bùguò wǒ měi zhōu zhǐ gěi tā xǐ yīcì.

22. Stim_T3T3_foc2

A: 你要洗手还是洗脚啊？

A: nǐ yào xǐ shǒu háishì xǐ jiǎo a?

B: 洗脚。

B: xǐ jiǎo.

A: 洗脚？

A: xǐ jiǎo?

B: 对，我的脚刚刚踩到泥了。

B: duì, wǒ de jiǎo gānggāng cǎi dào ní le.

23. Stim_T3T4_foc1

A: 你是想煮菜还是炒菜？

A: nǐ shì xiǎng zhǔ cài háishì chǎo cài?

B: 煮菜。

B: zhǔ cài.

A: 煮菜？

A: zhǔ cài?

B: 对，以前老炒菜，但后来发现煮菜更健康一些。

B: duì, yǐqián lǎo chǎo cài, dàn hòulái fāxiàn zhǔ cài gèng jiànkāng yīxiē.

24. Stim_T3T4_foc2

A: 你打算煮菜还是煮茶？

A: nǐ dǎsuàn zhǔ cài háishì zhǔ chá?

B: 煮菜。

B: zhǔ cài.

A: 煮菜?

A: zhǔ cài?

B: 对啊，难道你想喝茶吗？

B: duì a, nándào nǐ xiǎng hē chá ma?

25. Stim_T4T1_foc1

A: 你要去图书馆借书还是还书？

A: nǐ yào qù túshūguǎn jièshū háishì huánshū?

B: 借书。

B: jiè shū.

A: 借书？

A: jiè shū?

B: 对，老师叫我去借一本余华的老书。

B: duì, lǎoshī jiào wǒ qù jiè yī běn Yú Huá de lǎo shū.

26. Stim_T4T1_foc2

A: 你找我借书还是借电脑啊？

A: nǐ zhǎo wǒ jiè shū háishì jiè diànnǎo a?

B: 借书。

B: jiè shū.

A: 借书？

A: jiè shū?

B: 对，放心，我不是要跟你借电脑。

B: Duì, fàngxīn, bùshì gēn nǐ jiè diànnǎo.

27. Stim_T4T2_foc1

A: 他让你负责带茶还是泡茶？

A: tā ràng nǐ fùzé dài chá háishì pào chá?

B: 带茶。

B: dài chá.

A: 带茶？

A: dài chá?

B: 对，他来泡茶。

B: duì, tā lái pào chá.

28. Stim_T4T2_foc2

A: 你从云南回来，打算给你妈带茶还是带咖啡？

A: nǐ cóng Yúnnán huílái, dǎsuàn gěi nǐ mā dài chá háishì dài kāfēi?

B: 带茶。

B: dài chá.

A: 带茶？

A: dài chá?

B: 对啊，毕竟茶叶才是云南真正的特产。

B: duì a, bìjìng cháyè cái shì Yúnnán zhēnzhèng de tèchǎn.

29. Stim_T4T3_foc1

A: 他让你带酒还是品酒？

A: tā ràng nǐ dài jiǔ háishì pǐn jiǔ?

B: 带酒。

B: dài jiǔ.

A: 带酒？

A: dài jiǔ?

B: 对，他来品酒。

B: duì, tā lái pǐn jiǔ.

30. Stim_T4T3_foc2

A: 我去见岳父，带酒还是带水果呢？

A: wǒ qù jiàn yuèfù, dài jiǔ háishì dài shuǐguǒ ne?

B: 带酒。

B: dài jiǔ.

A: 带酒？

A: dài jiǔ?

B: 是啊，带水果多寒酸。

B: shì a, dài shuǐguǒ duō hánsuān.

31. Stim_T4T4_foc1

A: 你想做菜还是外卖点菜？

A: nǐ xiǎng zuò cài háishì wàimài diǎn cài?

B: 做菜。

B: zuò cài.

A: 做菜？

A: zuò cài?

B: 对，虽然有点麻烦，但最近总吃外卖，吃腻了。

B: duì, suīrán yǒudiǎn máfan, dàn zuìjìn lǎo diǎn cài, chī nì le.

32. Stim_T4T4_foc2

A: 你要做菜还是做点心？

A: nǐ yào zuò cài háishì zuò diǎnxīn?

B: 做菜。

B: zuò cài.

A: 做菜？

A: zuò cài?

B: 对，做点心的话时间太长。

B: duì, zuò diǎnxīn dehuà shíjiān tài cháng.

Appendix C: Mandarin Learning Background

Studio e pratica della lingua cinese

In questa sezione sono richieste alcune informazioni relative al proprio percorso di studio della lingua cinese.

41. Università di appartenenza *

Contrassegna solo un ovale.

- ◯ Università Roma Tre
- ◯ Università per Stranieri di Siena
- ◯ Università degli Studi di Perugia
- ◯ La Sapienza Università di Roma
- ◯ Università degli Studi di Napoli L'Orientale
- ◯ Altro: _____

42. Anno di corso *

Se fuori corso, specificare nella casella "Altro" indicando anche l'annualità dell'esame di cinese che si intende sostenere: ad es. "Fuori corso da un anno - Terzo (triennale)"

Contrassegna solo un ovale.

- ◯ Secondo (triennale)
- ◯ Terzo (triennale)
- ◯ Primo (magistrale)
- ◯ Secondo (magistrale)
- ◯ Altro: _____

Appendix Figure 4 Background questionnaire - Mandarin Learning Background section, p.1

43. Ore settimanali di lettorato universitario (lezioni con docente madrelingua) frequentate *

Rispondere in base alla media di ore settimanali che effettivamente si frequentano, non in base alle ore previste dal proprio corso di laurea

Contrassegna solo un ovale.

- ◯ 0 (fuori corso o non frequentante)
- ◯ 2
- ◯ 4
- ◯ 6
- ◯ Altro: _____

44. Considerando il metodo e l'approccio finora adottati nello studio della lingua cinese, valutare su una scala da 1 (=pochissima) a 5 (=moltissima) l'importanza che si attribuisce alle seguenti abilità linguistiche *

Non basare la risposta su ciò che si avrebbe voluto o che si vorrebbe fare, ma su ciò che effettivamente si ritiene di fare

Contrassegna solo un ovale per riga.

	1 (Pochissima importanza)	2	3	4	5 (Moltissima importanza)
Grammatica	◯	◯	◯	◯	◯
Scrittura	◯	◯	◯	◯	◯
Vocabolario	◯	◯	◯	◯	◯
Parlato	◯	◯	◯	◯	◯

Appendix Figure 5 Background questionnaire - Mandarin Learning Background section, p.2

45. Sono interessato ad acquisire una pronuncia accurata in cinese *

1=Pochissimo; 5=Moltissimo

Contrassegna solo un ovale.

	1	2	3	4	5	
Pocl	◯	◯	◯	◯	◯	Moltissimo

46. Svolgo esercizi specifici mirati al perfezionamento della mia pronuncia in cinese *

Puoi scegliere più di un'opzione affermativa

Seleziona tutte le voci applicabili.

- ☐ No
- ☐ Sì, all'università
- ☐ Sì, durante ripetizioni o corsi di lingua non curricolari (ad es. Istituto Confucio, scuole private, ecc.)
- ☐ Sì, in autonomia

47. Durante le lezioni di lettorato utilizzo oralmente la lingua cinese per esercitarmi e comunicare *

1=Pochissimo; 5=Moltissimo

Contrassegna solo un ovale.

	1	2	3	4	5	
Pocl	◯	◯	◯	◯	◯	Moltissimo

Appendix Figure 6 Background questionnaire - Mandarin Learning Background section, p.3

48. Di solito, uso la lingua cinese anche al di fuori di contesti didattici *

Ad es., per lavoro, con amici cinesi, ecc.

Contrassegna solo un ovale.

	1	2	3	4	5	
Per	◯	◯	◯	◯	◯	Molto spesso

49. Trovo che lo studio della lingua cinese sia... *

Puoi scegliere più di un'opzione

Seleziona tutte le voci applicabili.

- ☐ Facile
- ☐ Difficile
- ☐ Divertente
- ☐ Noioso
- ☐ Stimolante
- ☐ Demotivante
- ☐ Altro: _____

Appendix Figure 7 Background questionnaire - Mandarin Learning Background section, p.4

Appendix D: Sociolinguistic and Biographical Background

50. Età *

Contrassegna solo un ovale.

- ◯ 20
- ◯ 21
- ◯ 22
- ◯ 23
- ◯ 24
- ◯ 25
- ◯ Altro: _____

51. In quale città vivi attualmente? *

Se si tratta di un comune, inserire la provincia tra parentesi, ad es. "Velletri (Roma)"

_____

52. Attualmente vivi nella stessa città in cui sei cresciuto/a? *

Contrassegna solo un ovale.

- ◯ Sì
- ◯ No

53. Se no, specificare indicativamente da che età a che età hai soggiornato in ciascuna città *

Ad es. "Roma: 0-3 anni; Parigi: 4-10 anni; Torino: 11-26 anni"
Includere soltanto i soggiorni più significativi, non inferiori a due anni

_____

Appendix Figure 8 Background questionnaire - Sociolinguistic and Biographical Background section, p.1

54. I tuoi genitori sono entrambi di madrelingua italiana? *

Contrassegna solo un ovale.

- ◯ Sì
- ◯ No

55. Se no, specificare

_____

56. Sei bi/plurilingue? *

ℹ *In questo caso, "bi/plurilingue" si riferisce alla condizione di essere nati e cresciuti parlando una o più lingue oltre all'italiano, in modo più o meno equilibrato*

Contrassegna solo un ovale.

- ◯ Sì
- ◯ No

57. Se sì, in quale/i lingua/i oltre all'italiano?

_____

58. Prima dell'università, hai frequentato lezioni di lingua cinese?

Puoi scegliere più di un'opzione affermativa

Seleziona tutte le voci applicabili.

- ☐ No
- ☐ Sì, alla scuola superiore (insegnamento curricolare della lingua cinese)
- ☐ Sì, alla scuola superiore (insegnamento non curricolare della lingua cinese)
- ☐ Sì, presso un Istituto Confucio
- ☐ Sì, ho fatto ripetizioni private
- ☐ Altro: _____

Appendix Figure 9 Background questionnaire - Sociolinguistic and Biographical Background section, p.2

59. Se sì, per quanto tempo?

Contrassegna solo un ovale.

- ◯ Meno di 6 mesi
- ◯ Da 6 mesi a 1 anno
- ◯ 1-2 anni
- ◯ 3-4 anni
- ◯ 5 anni o più

60. Durante l'università, hai frequentato/frequenti lezioni non curricolari di lingua cinese? *

Puoi scegliere più di un'opzione affermativa

Seleziona tutte le voci applicabili.

- ☐ No
- ☐ Sì, presso un Istituto Confucio
- ☐ Sì, lezioni private
- ☐ Altro: _____

61. Se sì, per/da quanto tempo?

Contrassegna solo un ovale.

- ◯ Meno di 6 mesi
- ◯ Da 6 mesi a 1 anno
- ◯ 1-2 anni
- ◯ 3-4 anni
- ◯ 5 anni o più

Appendix Figure 10 Background questionnaire - Sociolinguistic and Biographical Background section, p.2

62. Sei mai stato/a in Cina? *

Contrassegna solo un ovale.

- ◯ Sì
- ◯ No

63. Se sì, per quale motivo?

Puoi scegliere più di un'opzione

Seleziona tutte le voci applicabili.

- ☐ Mobilità universitaria
- ☐ Viaggio studio
- ☐ Lavoro
- ☐ Au Pair (ragazza/o alla pari)
- ☐ Altro: _____

64. Se sì, per quanto tempo in totale?

Contrassegna solo un ovale.

- ◯ Meno di 6 mesi
- ◯ Da 6 mesi a 1 anno
- ◯ 1-2 anni
- ◯ 3-4 anni
- ◯ 5 anni o più

Appendix Figure 11 Background questionnaire - Sociolinguistic and Biographical Background section, p.3

Appendix E: Tone identification test score per speaker

Appendix Table 1 Tone identification test score per speaker

Speaker	Monosyllabic target accuracy score (%)	Disyllabic target accuracy score (%)	Overall identification score (%)
S1	68.75	31.25	50
S2	62.5	56.25	59.375
S3	43.75	46.88	45.315
S4	68.75	68.75	68.75
S5	81.25	59.38	70.315
S6	56.25	50	53.125
S7	62.5	34.38	48.44
S8	68.75	62.5	65.625
S9	50	50	50
S10	75	56.25	65.625
S11	68.75	65.63	67.19
S12	100	96.88	98.44
S13	93.75	78.13	85.94
S14	68.75	65.63	67.19
S15	81.25	40.63	60.94
S16	75	31.25	53.125
S17	37.5	37.5	37.5
S18	56.25	59.38	57.815
S19	75	65.63	70.315
S20	75	65.63	70.315
S21	56.25	40.63	48.44
S22	75	71.88	73.44
S23	75	62.5	68.75
S24	75	43.75	59.375
S25	62.5	50	56.25
S26	93.75	62.5	78.125
S27	62.5	40.63	51.565
S28	100	96.88	98.44
S29	81.25	34.38	57.815
S30	68.75	50	59.375
S31	81.25	43.75	62.5
S32	56.25	50	53.125
S33	68.75	40.63	54.69

S34	87.5	84.38	85.94
S35	87.5	43.75	65.625
S36	87.5	43.75	65.625
S37	31.25	28.13	29.69
S38	93.75	84.38	89.065
S39	81.25	68.75	75
S40	62.5	46.88	54.69
S41	75	59.38	67.19
S42	87.5	50	68.75

Appendix F: Proficiency score clustering

Appendix Table 2 Proficiency score clustering

Speaker	HSKK_score	HSKK_z	Id_score	Id_z	K2Proficiency	K3Proficiency
S1	11.5	-0.97818	50	-0.93551	Low	Low
S2	21	0.543967	59.375	-0.28791	Low	Low
S3	9.5	-1.29863	45.315	-1.25914	Low	Low
S4	14.5	-0.4975	68.75	0.359687	Low	Intermediate
S5	9.75	-1.25857	70.315	0.467792	Low	Intermediate
S6	13.25	-0.69778	53.125	-0.71964	Low	Low
S7	15.33	-0.36451	48.44	-1.04327	Low	Low
S8	20.5	0.463854	65.625	0.14382	Low	Intermediate
S9	12.25	-0.85801	50	-0.93551	Low	Low
S10	11.25	-1.01824	65.625	0.14382	Low	Intermediate
S11	11	-1.05829	67.19	0.251926	Low	Intermediate
S12	13.75	-0.61767	98.44	2.410589	High	Intermediate
S13	12.5	-0.81795	85.94	1.547124	Low	Intermediate
S14	13.25	-0.69778	67.19	0.251926	Low	Intermediate
S15	20.67	0.491092	60.94	-0.17981	Low	Low
S16	13.25	-0.69778	53.125	-0.71964	Low	Low
S17	11.5	-0.97818	37.5	-1.79898	Low	Low
S18	16	-0.25716	57.815	-0.39567	Low	Low
S19	15.5	-0.33728	70.315	0.467792	Low	Intermediate
S20	12.75	-0.7779	70.315	0.467792	Low	Intermediate
S21	9.75	-1.25857	48.44	-1.04327	Low	Low
S22	26.33	1.397971	73.44	0.683659	High	High
S23	19.33	0.27639	68.75	0.359687	Low	Intermediate
S24	13	-0.73784	59.375	-0.28791	Low	Low
S25	12.25	-0.85801	56.25	-0.50378	Low	Low
S26	12.25	-0.85801	78.125	1.007285	Low	Intermediate
S27	17.5	-0.01682	51.565	-0.82741	Low	Low
S28	26.5	1.425209	98.44	2.410589	High	High
S29	17.5	-0.01682	57.815	-0.39567	Low	Low
S30	19.75	0.343685	59.375	-0.28791	Low	Low
S31	30.25	2.026056	62.5	-0.07205	High	High
S32	26.25	1.385153	53.125	-0.71964	High	High
S33	20.75	0.50391	54.69	-0.61154	Low	Low
S34	27.5	1.585435	85.94	1.547124	High	High
S35	19.5	0.303628	65.625	0.14382	Low	Intermediate
S36	25.25	1.224927	65.625	0.14382	High	High
S37	16	-0.25716	29.69	-2.33847	Low	Low
S38	31.25	2.186282	89.065	1.76299	High	High

S39	29.75	1.945943	75	0.791419	High	High
S40	19.75	0.343685	54.69	-0.61154	Low	Low
S41	25	1.18487	67.19	0.251926	High	High
S42	15	-0.41739	68.75	0.359687	Low	Intermediate

Appendix G: Musicality score clustering

Appendix Table 3 Musicality score loadings per speaker

Speaker	MSI_score	MSI_z	Td_score	Td_z	Musicality_zmean	Musicality_pca
S1	44	-1.4508614	28	0.67494856	-0.387956399	-0.509006279
S2	64	-0.1529662	28	0.67494856	0.260991172	0.342425453
S3	55	-0.737019	28	0.67494856	-0.031035235	-0.040718827
S4	68	0.10661282	29	1.01242284	0.559517826	0.734098181
S5	72	0.36619184	26	0	0.183095922	0.240225382
S6	82	1.01513942	28	0.67494856	0.845043986	1.108714011
S7	63	-0.217861	27	0.33747428	0.059806654	0.078467484
S8	94	1.7938765	26	0	0.89693825	1.176800287
S9	79	0.82045514	23	-1.0124228	-0.095983846	-0.125932658
S10	63	-0.217861	28	0.67494856	0.228543794	0.299853866
S11	69	0.17150757	31	1.68737139	0.929439483	1.219442531
S12	67	0.04171806	25	-0.3374743	-0.14787811	-0.194018933
S13	75	0.56087612	28	0.67494856	0.617912336	0.810712905
S14	52	-0.9317033	25	-0.3374743	-0.634588789	-0.832592732
S15	61	-0.3476505	29	1.01242284	0.332386176	0.436097075
S16	50	-1.0614928	22	-1.3498971	-1.205694964	-1.581895051
S17	75	0.56087612	28	0.67494856	0.617912336	0.810712905
S18	62	-0.2827557	27	0.33747428	0.027359276	0.035895897
S19	78	0.75556039	25	-0.3374743	0.209043054	0.274268519
S20	104	2.44282407	28	0.67494856	1.558886315	2.045288916
S21	66	-0.0231767	17	-3.0372685	-1.530222604	-2.007681575
S22	85	1.20982369	30	1.34989712	1.279860401	1.679201535
S23	38	-1.8402299	25	-0.3374743	-1.088852089	-1.428594945
S24	109	2.76729786	24	-0.6749486	1.04617465	1.372601322
S25	58	-0.5423348	24	-0.6749486	-0.608641657	-0.798549595
S26	48	-1.1912823	28	0.67494856	-0.258166885	-0.338719933
S27	46	-1.3210718	27	0.33747428	-0.491798781	-0.645249488
S28	71	0.30129709	28	0.67494856	0.488122822	0.640426559
S29	46	-1.3210718	22	-1.3498971	-1.335484479	-1.752181398
S30	58	-0.5423348	21	-1.6873714	-1.114853075	-1.46270874
S31	65	-0.0880715	28	0.67494856	0.293438551	0.384997039
S32	73	0.4310866	27	0.33747428	0.38428044	0.50418335
S33	70	0.23640233	25	-0.3374743	-0.050535975	-0.066304174
S34	73	0.4310866	24	-0.6749486	-0.121930978	-0.159975796
S35	62	-0.2827557	22	-1.3498971	-0.816326421	-1.071036012
S36	48	-1.1912823	23	-1.0124228	-1.101852582	-1.445651843
S37	73	0.4310866	24	-0.6749486	-0.121930978	-0.159975796
S38	82	1.01513942	29	1.01242284	1.013781126	1.330100393

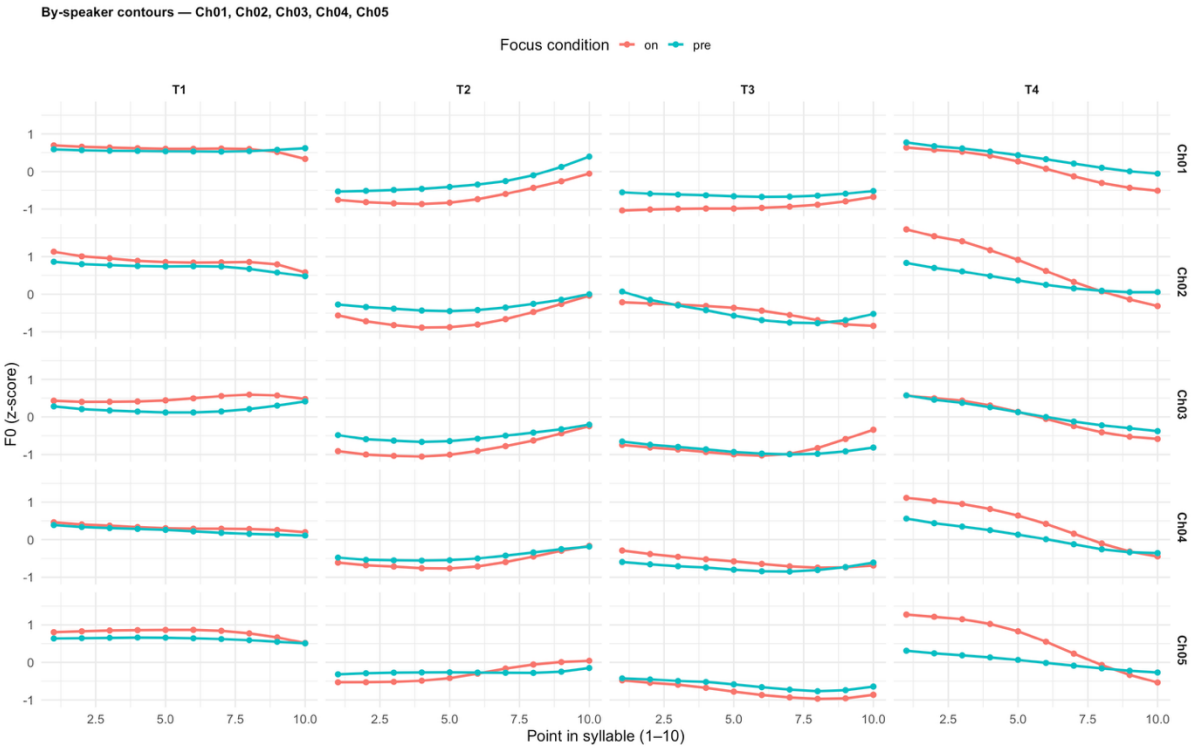
S39	60	-0.4125452	30	1.34989712	0.468675937	0.61491187
S40	66	-0.0231767	21	-1.6873714	-0.855274047	-1.122136048
S41	70	0.23640233	29	1.01242284	0.624412583	0.819241354
S42	43	-1.5157561	25	-0.3374743	-0.926615196	-1.215737012

Appendix Table 4 Musicality clustering per speaker

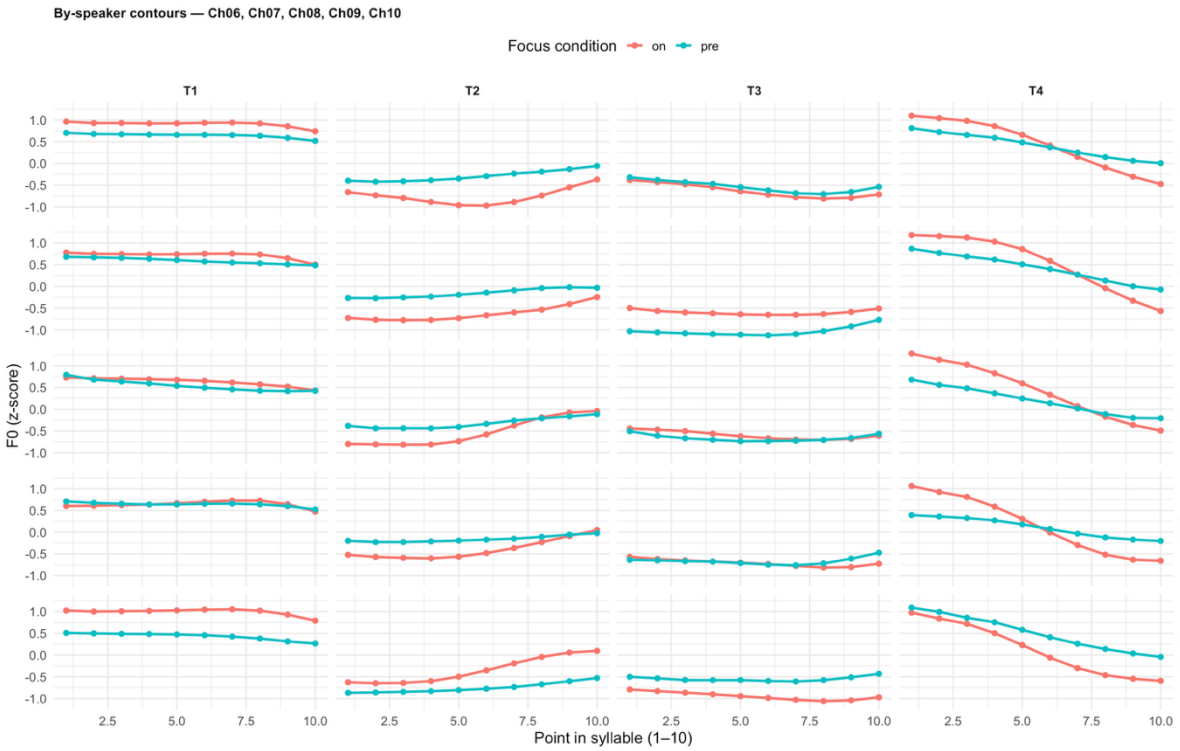
Speaker	**Musicality**
S1	Low
S2	High
S3	High
S4	High
S5	High
S6	High
S7	High
S8	High
S9	High
S10	High
S11	High
S12	High
S13	High
S14	Low
S15	High
S16	Low
S17	High
S18	High
S19	High
S20	High
S21	Low
S22	High
S23	Low
S24	High
S25	Low
S26	Low
S27	Low
S28	High
S29	Low
S30	Low
S31	High
S32	High
S33	High
S34	High

S35	Low
S36	Low
S37	High
S38	High
S39	High
S40	Low
S41	High
S42	Low

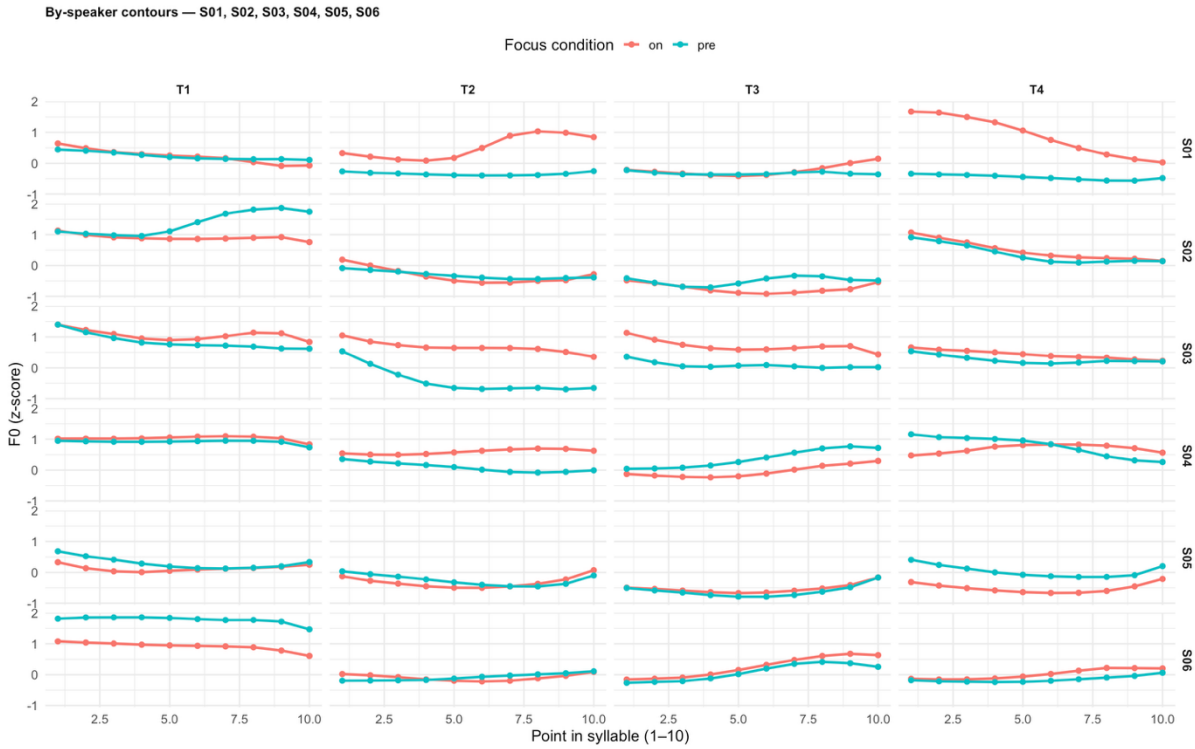
Appendix H: Focus analysis By-speaker contours



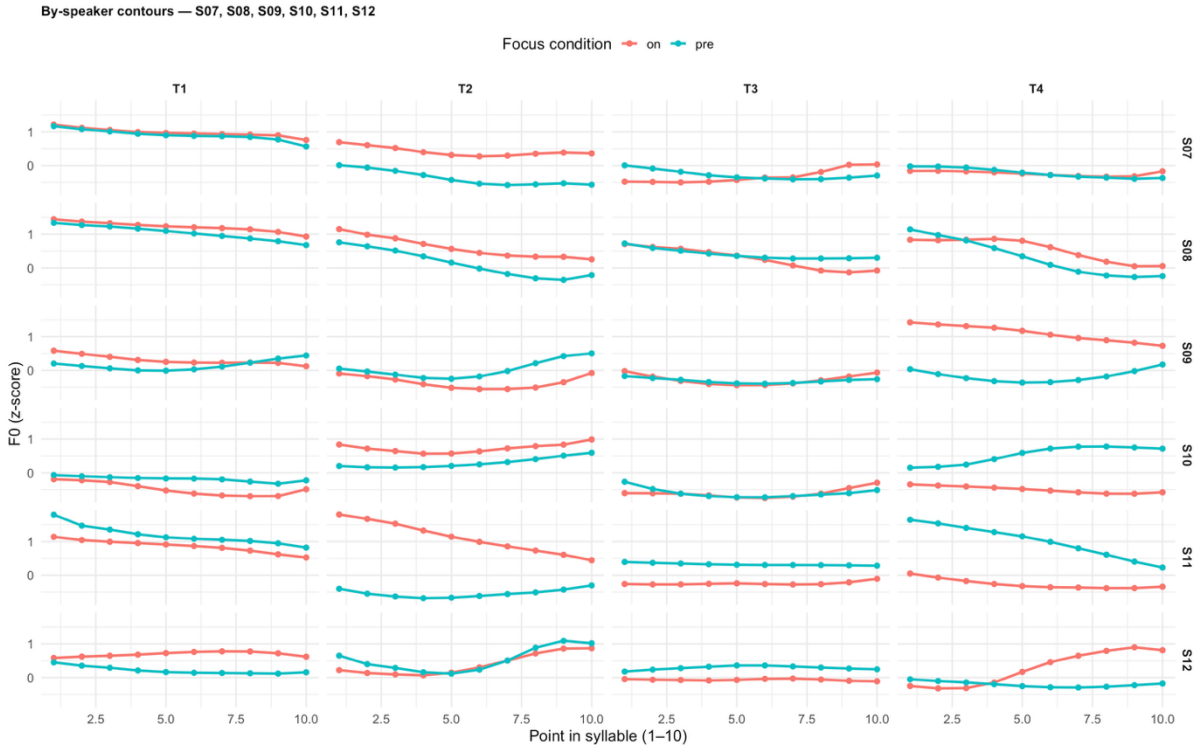
Appendix Figure 12 Focus analysis on Syl1, By-speaker contours (Ch01-Ch05)



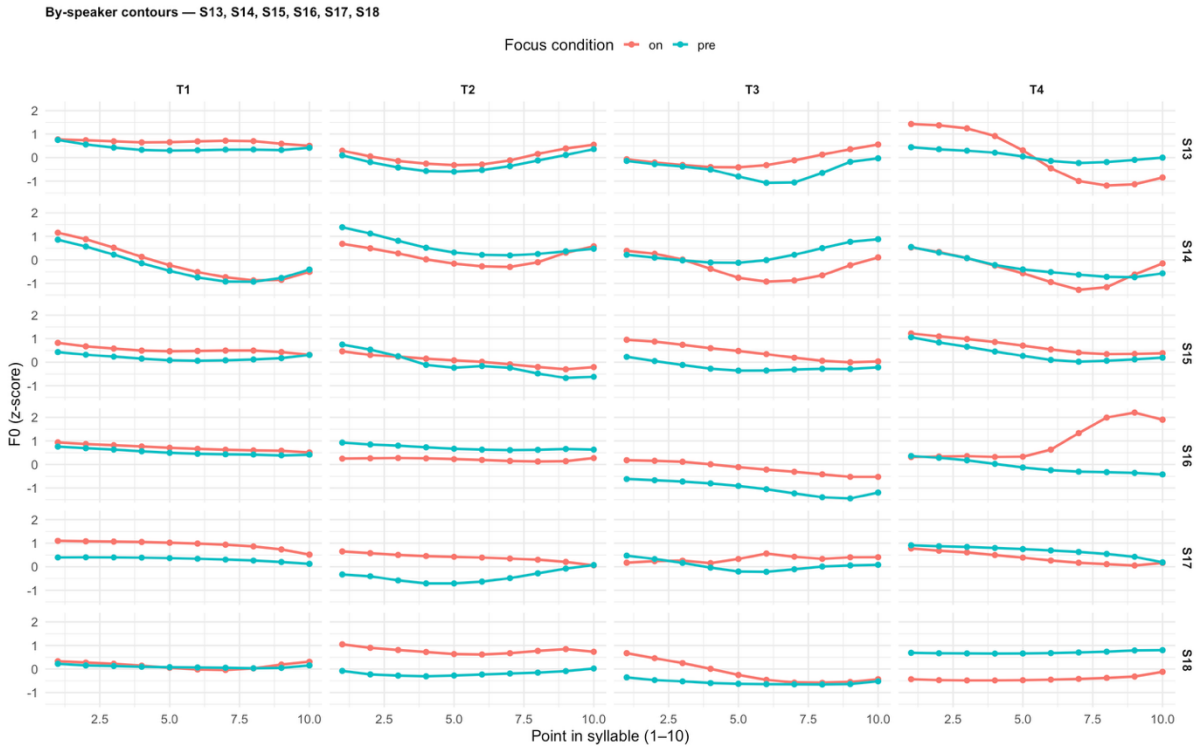
Appendix Figure 13 Focus analysis on Syl1, By-speaker contours (Ch06-Ch10)



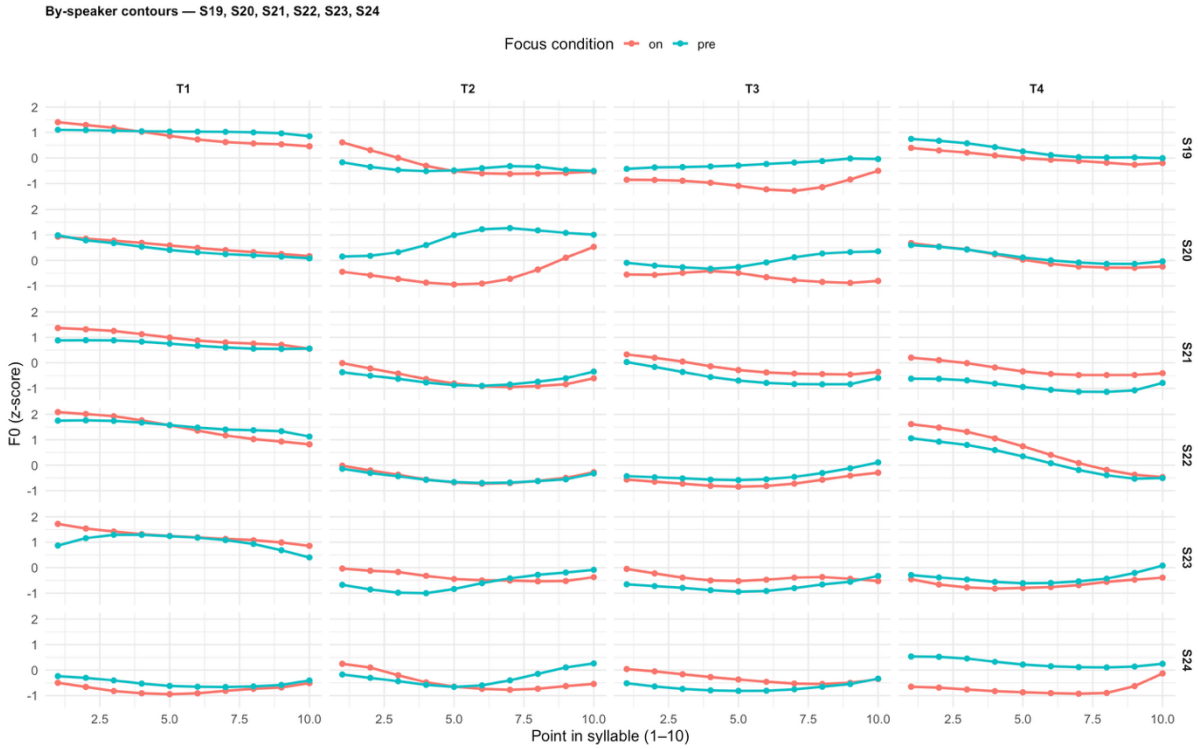
Appendix Figure 14 Focus analysis on Syl1, By-speaker contours (S01-S06)



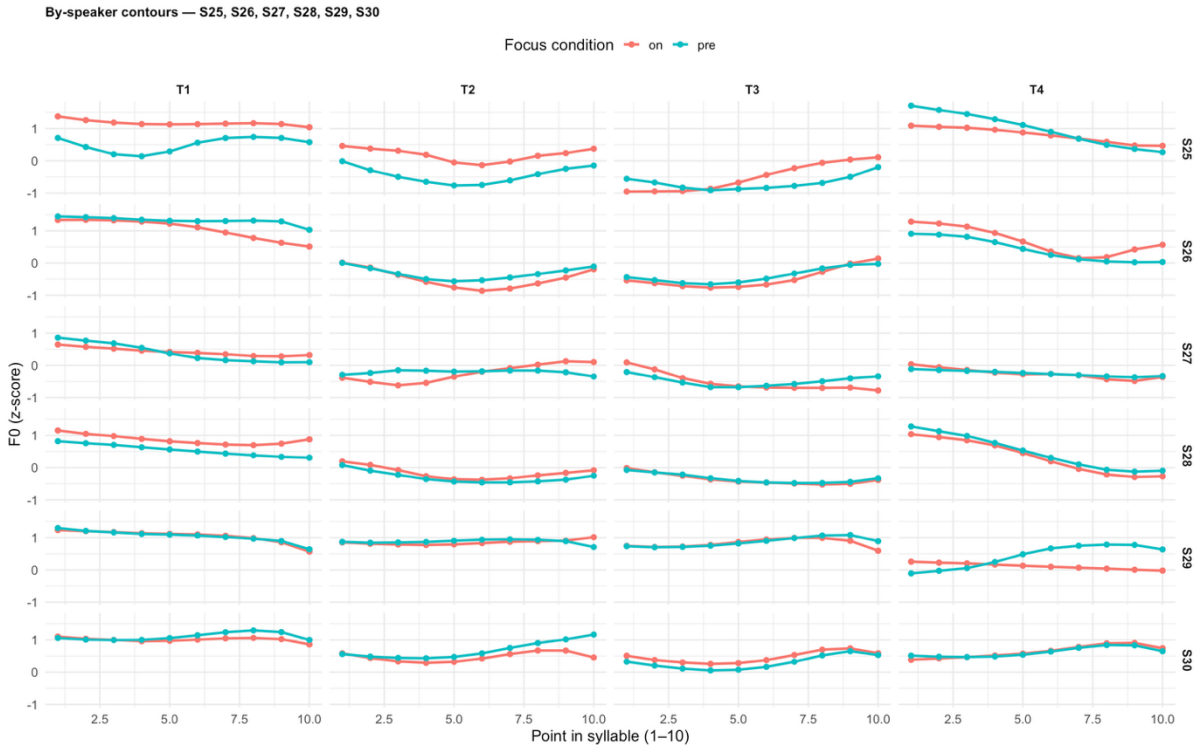
Appendix Figure 15 Focus analysis on Syl1, By-speaker contours (S07-S12)



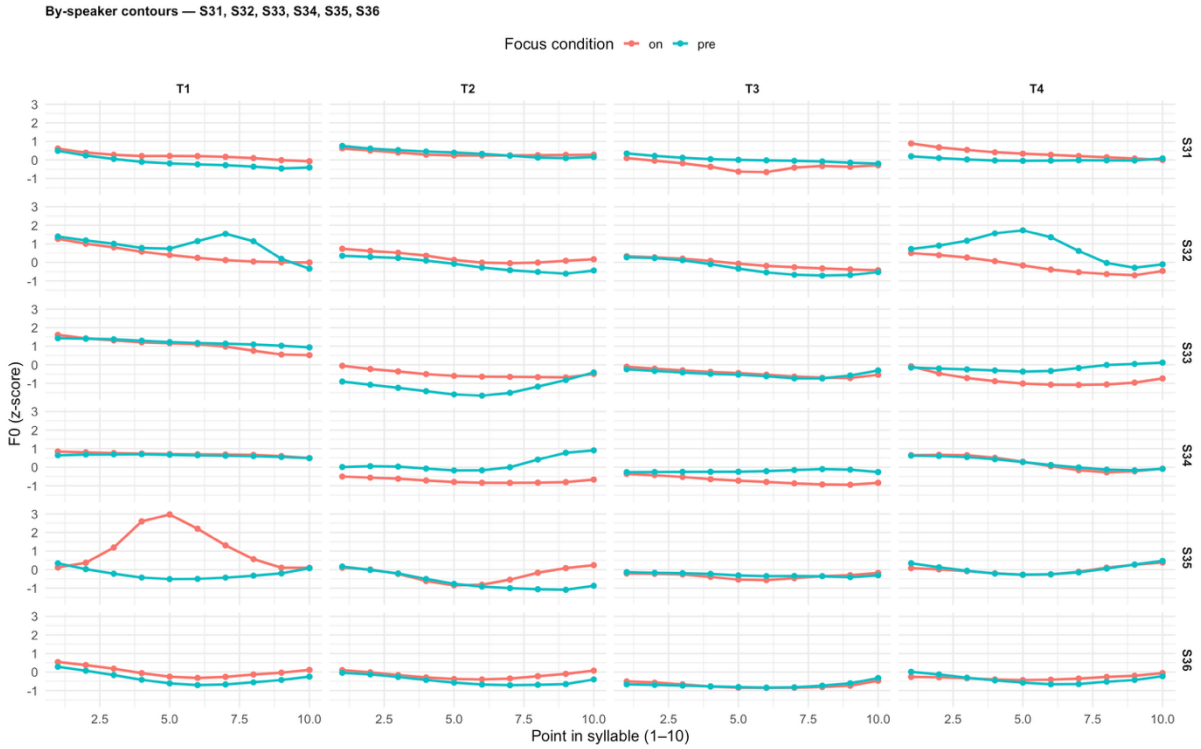
Appendix Figure 16 Focus analysis on Syl1, By-speaker contours (S13-S18)



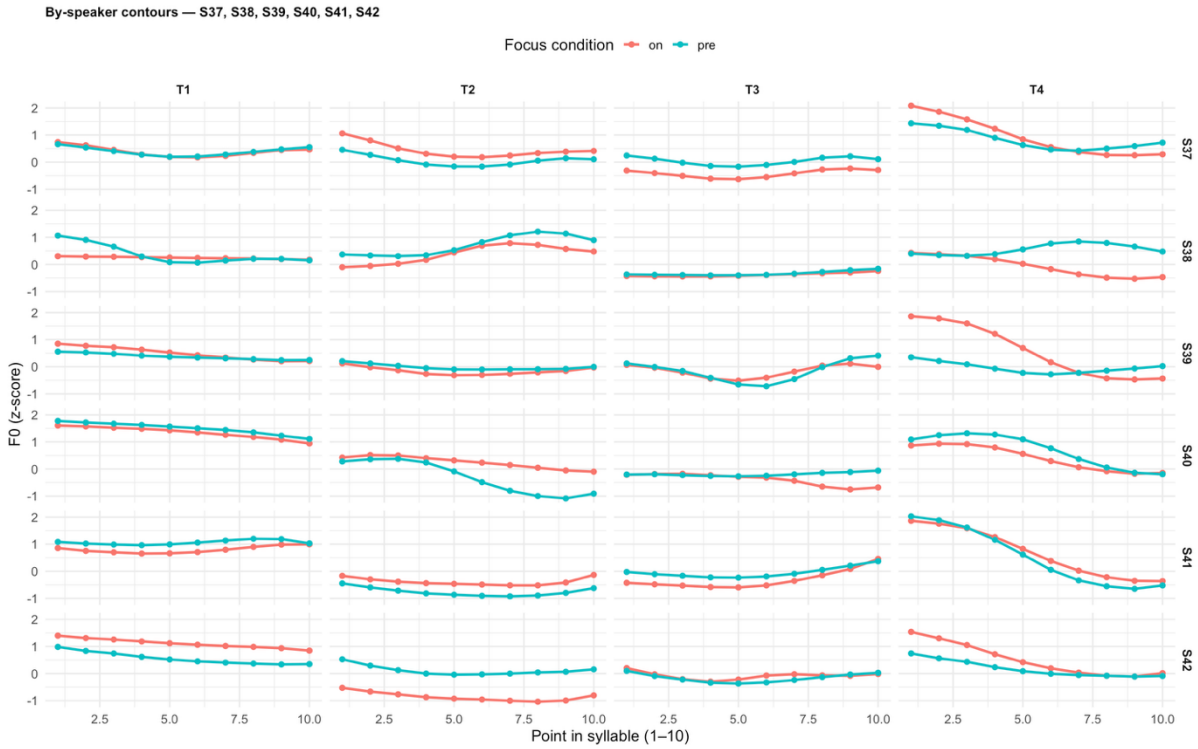
Appendix Figure 17 Focus analysis on Syl1, By-speaker contours (S19-S24)



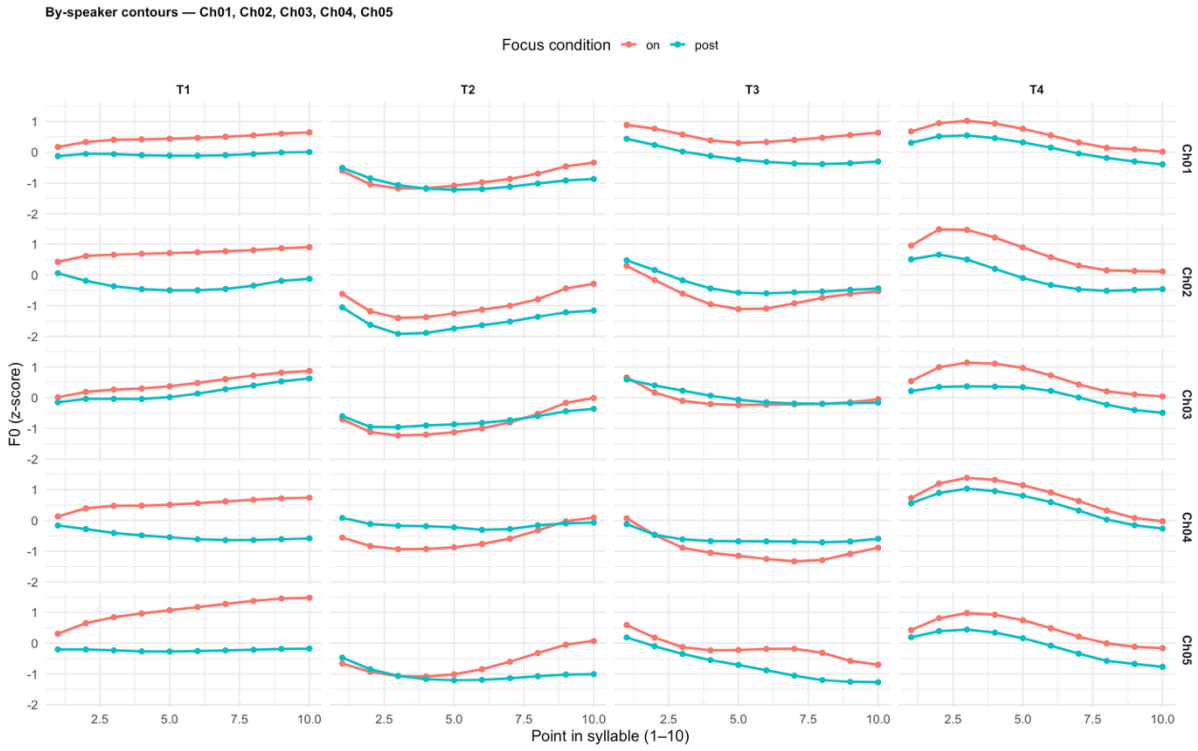
Appendix Figure 18 Focus analysis on Syl1, By-speaker contours (S25-S30)



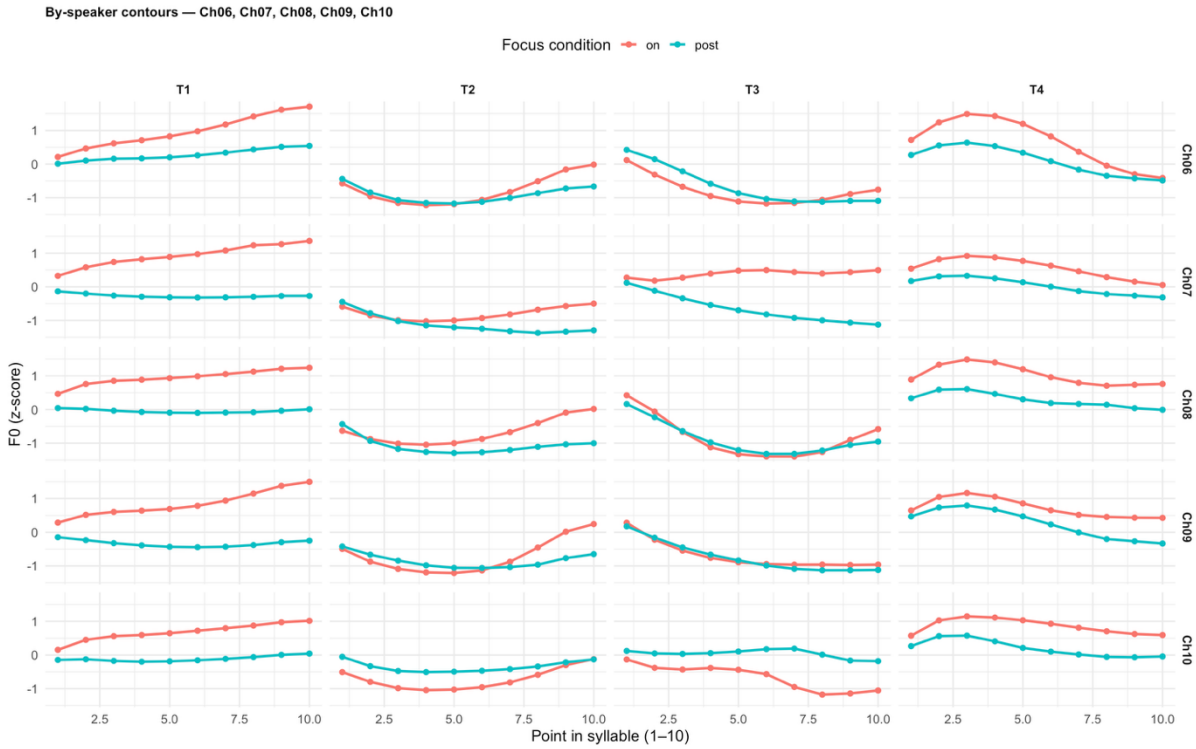
Appendix Figure 19 Focus analysis on Syl1, By-speaker contours (S31-S36)



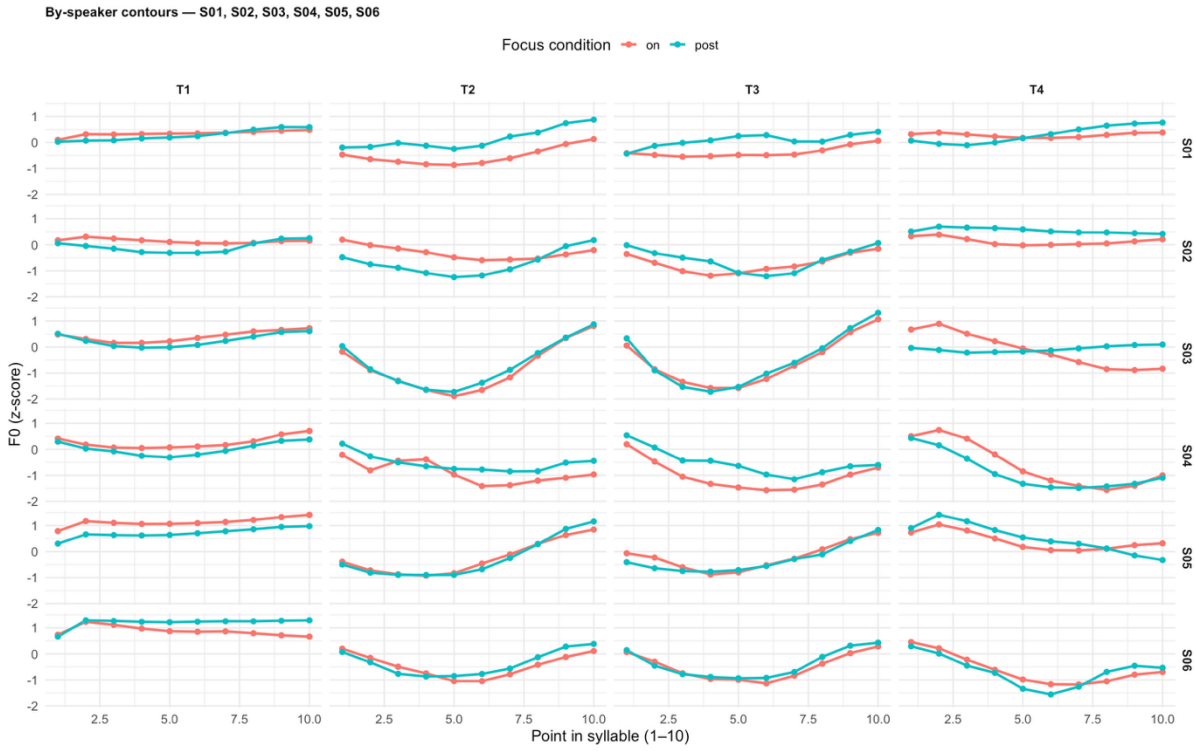
Appendix Figure 20 Focus analysis on Syl1, By-speaker contours (S37-S42)



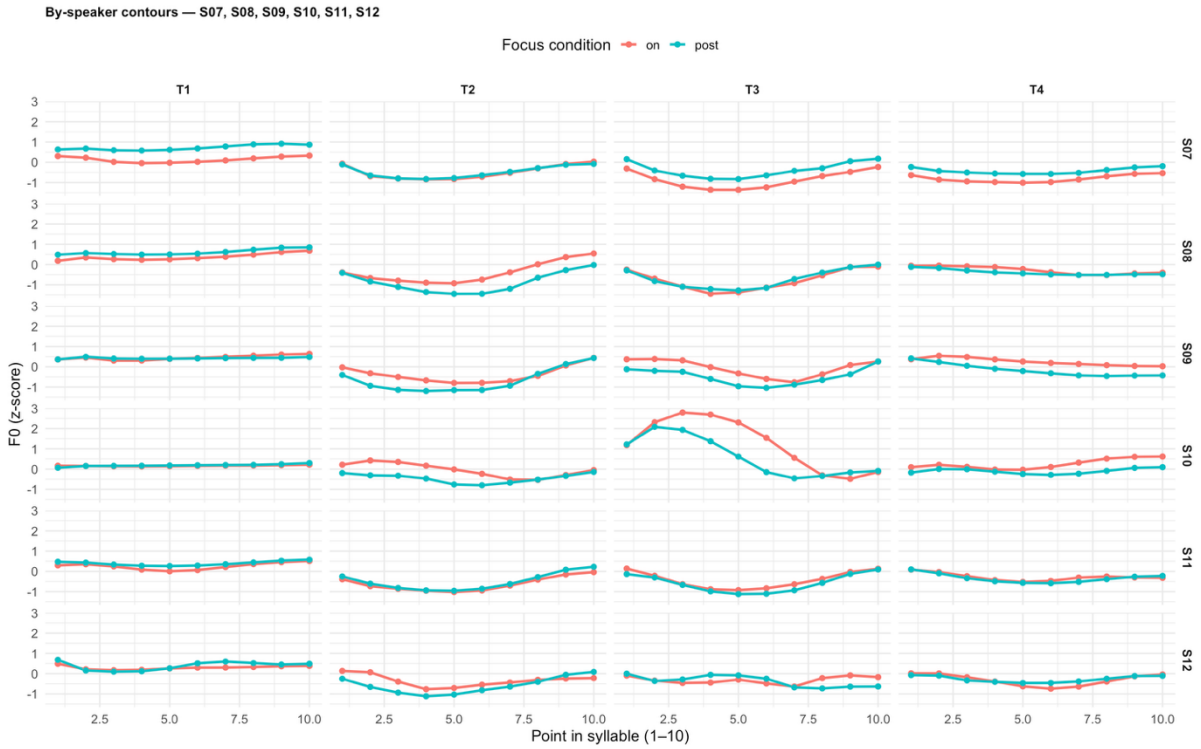
Appendix Figure 21 Focus analysis on Syl2, By-speaker contours (Ch01-Ch05)



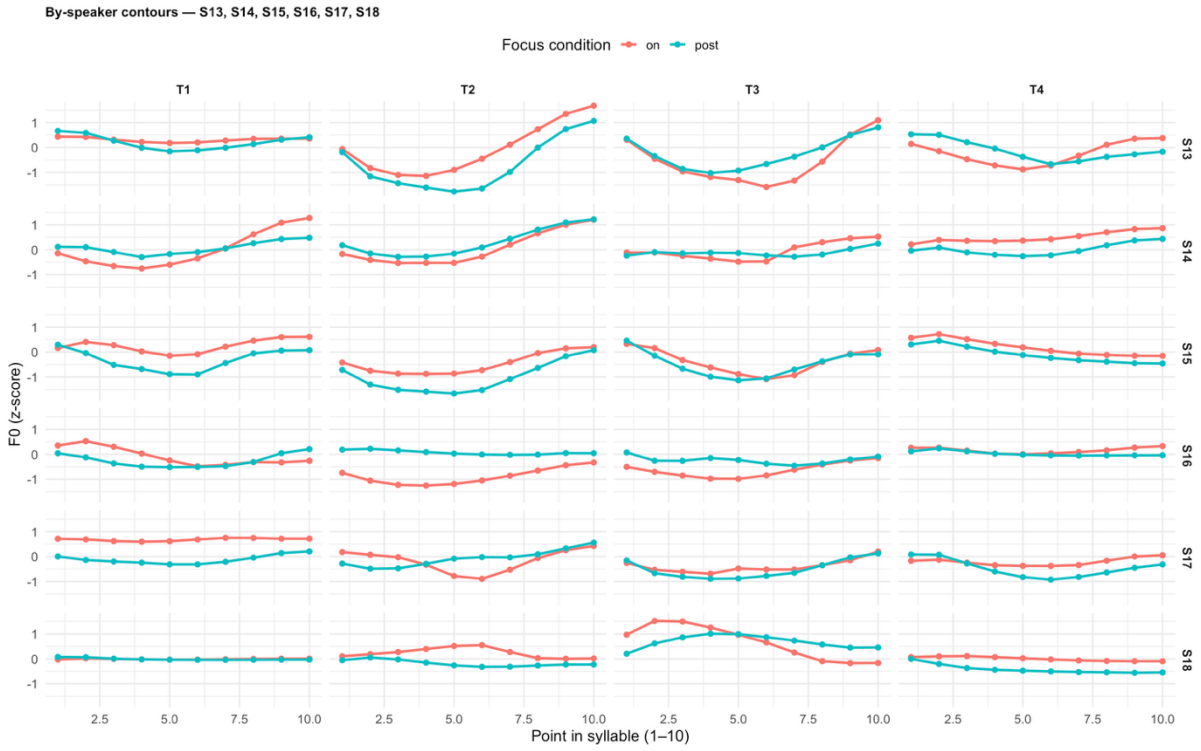
Appendix Figure 22 Focus analysis on Syl2, By-speaker contours (Ch06-Ch10)



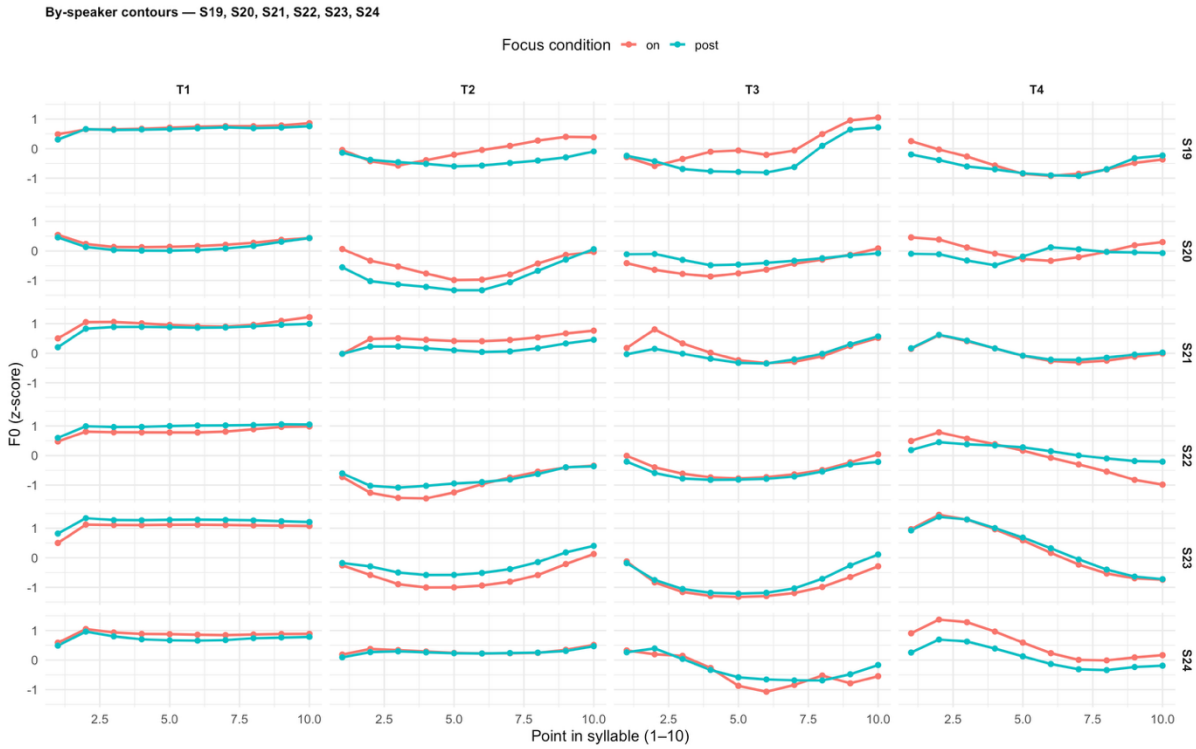
Appendix Figure 23 Focus analysis on Syl2, By-speaker contours (S01-S06)



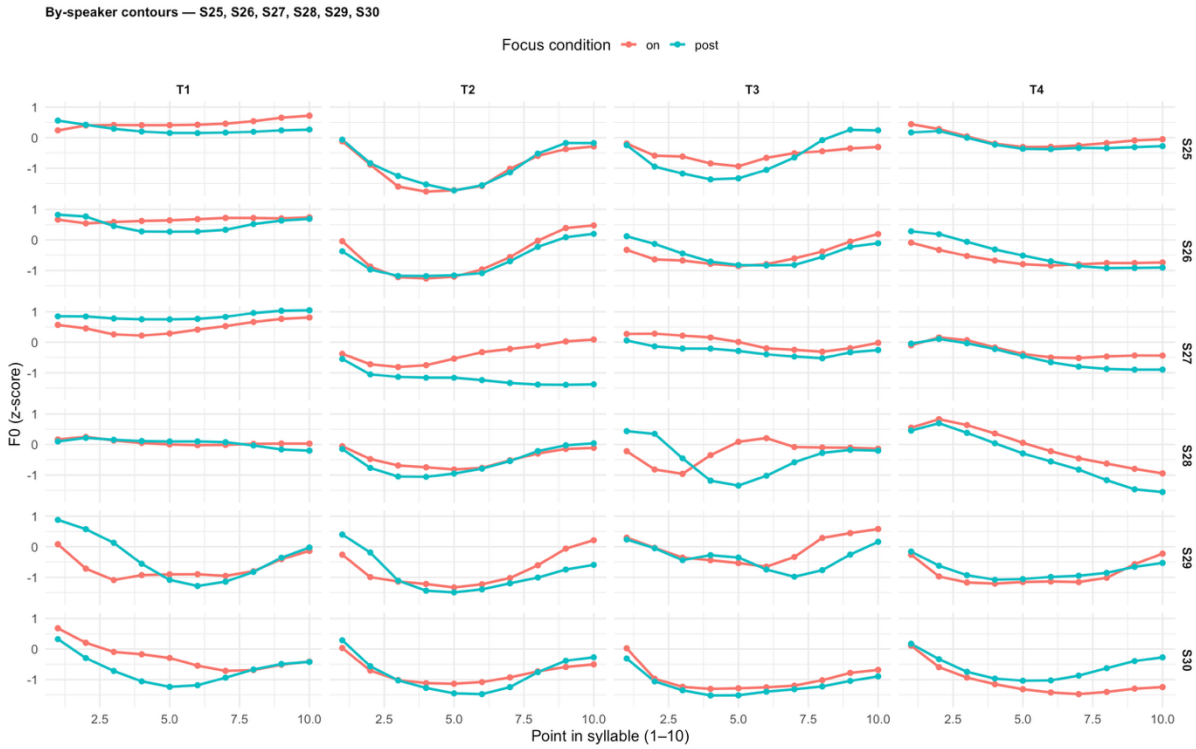
Appendix Figure 24 Focus analysis on Syl2, By-speaker contours (S07-S12)



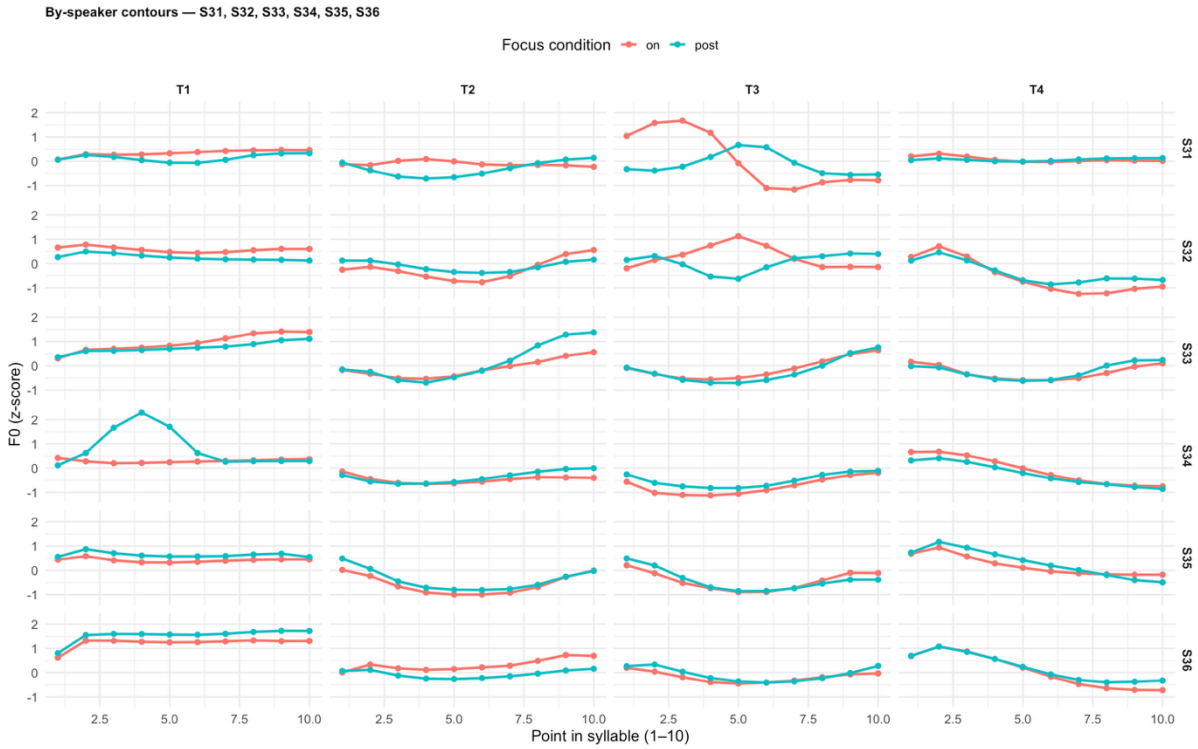
Appendix Figure 25 Focus analysis on Syl2, By-speaker contours (S013-S18)



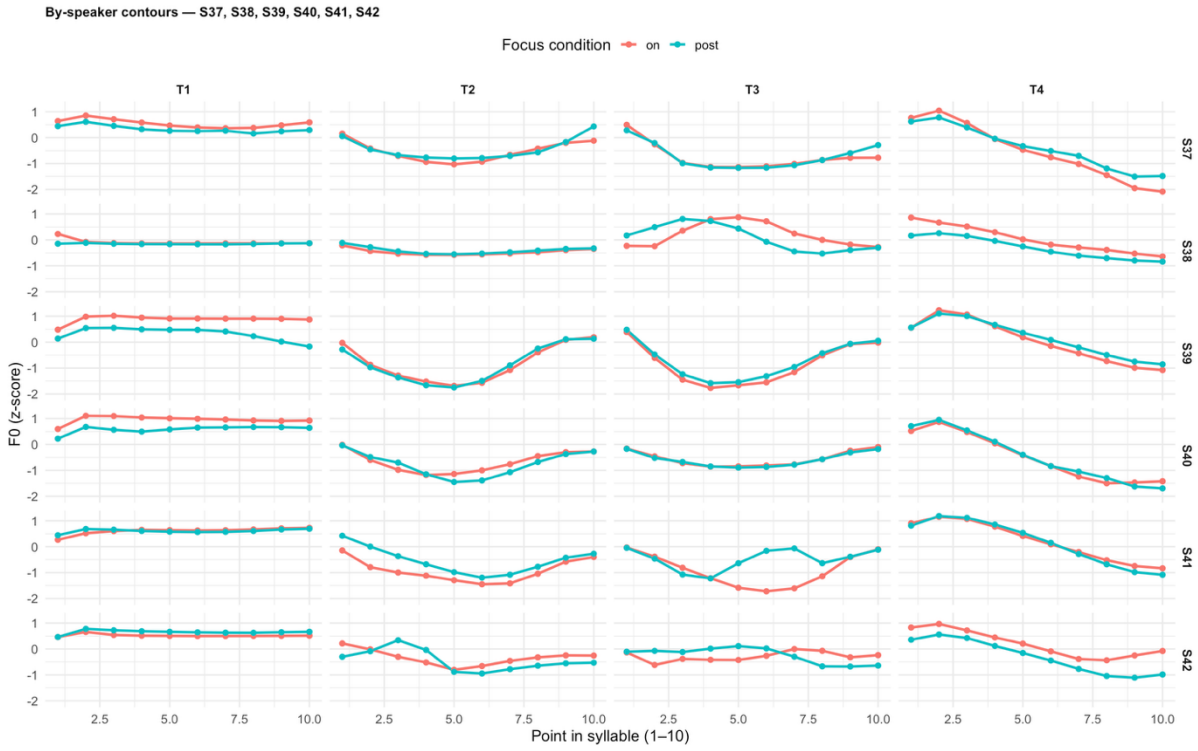
Appendix Figure 26 Focus analysis on Syl2, By-speaker contours (S19-S24)



Appendix Figure 27 Focus analysis on Syl2, By-speaker contours (S25-S30)

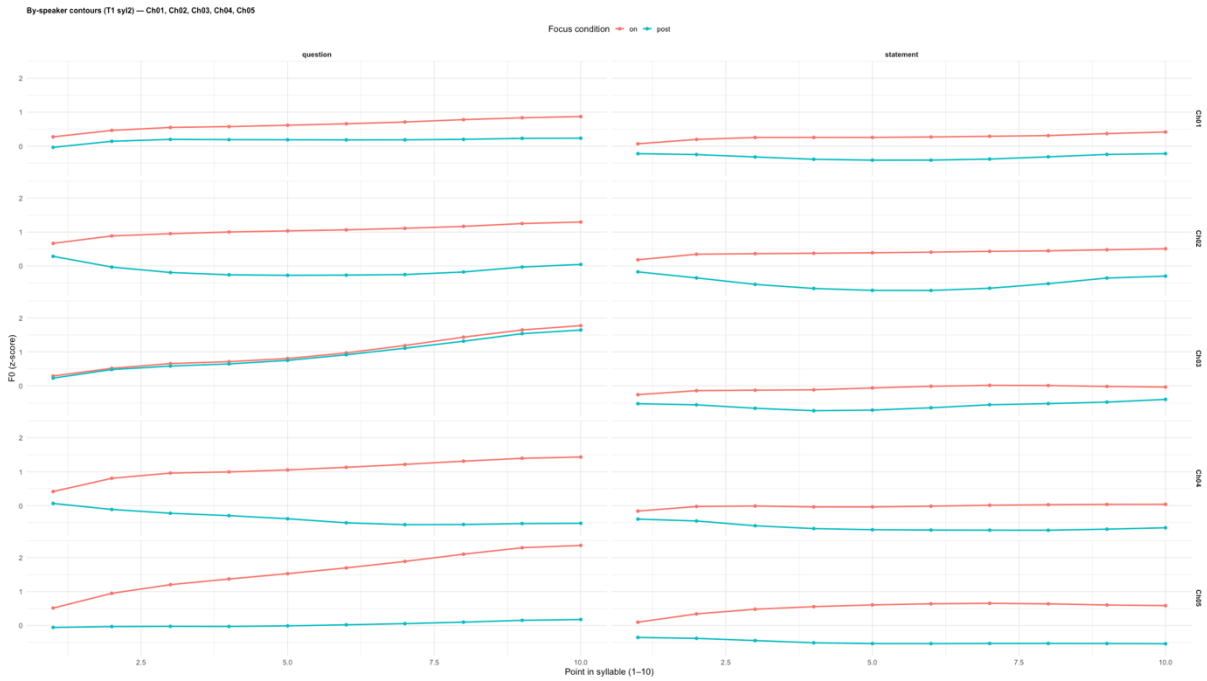


Appendix Figure 28 Focus analysis on Syl2, By-speaker contours (S31-S36)

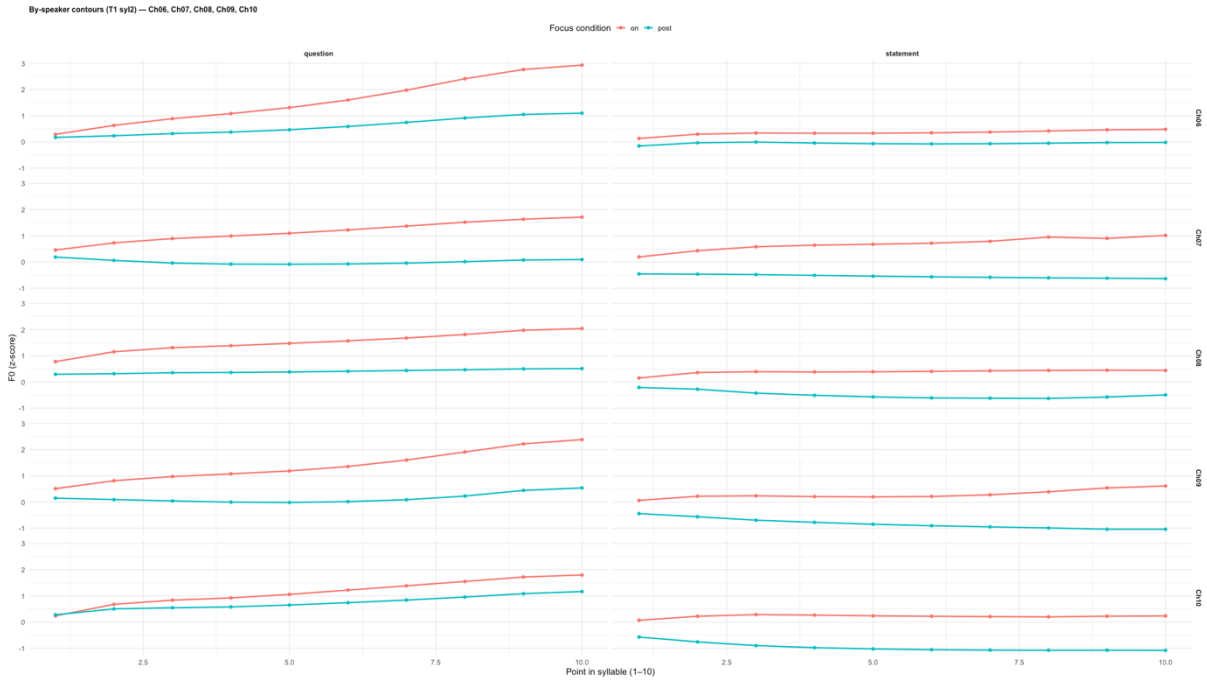


Appendix Figure 29 Focus analysis on Syl2, By-speaker contours (S37-S42)

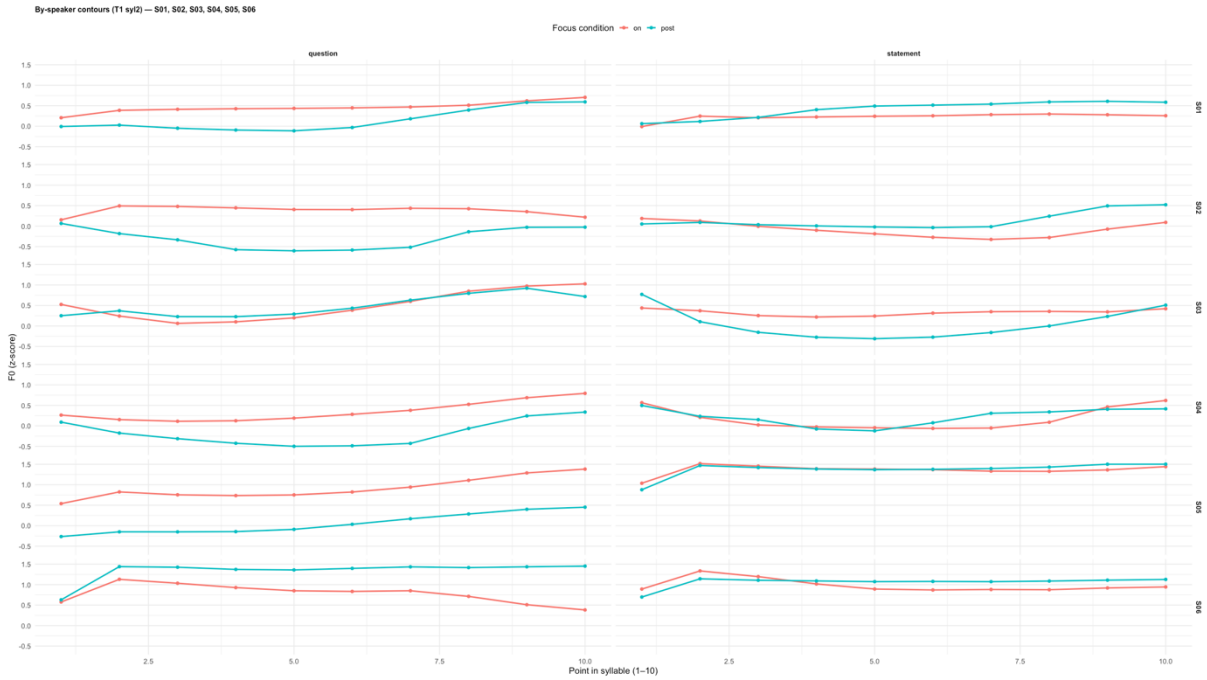
Appendix I: S.Type analysis By-speaker contours



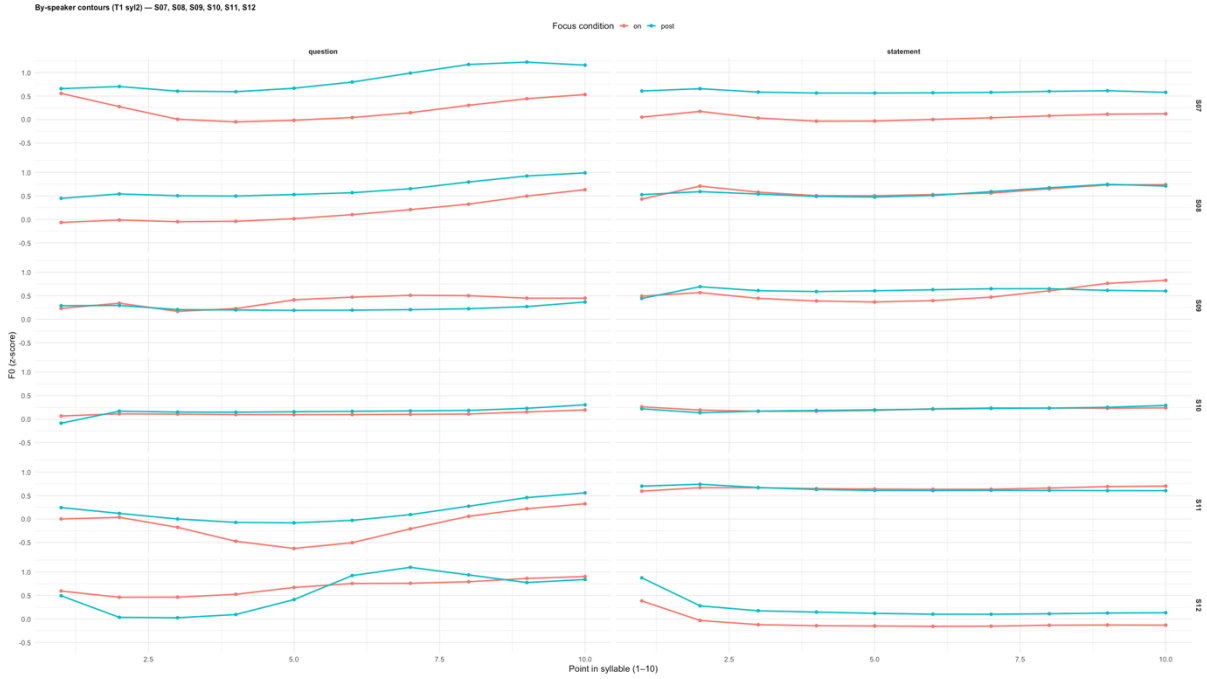
Appendix Figure 30 S.Type analysis on T1, By-speaker contours (Ch01-Ch05)



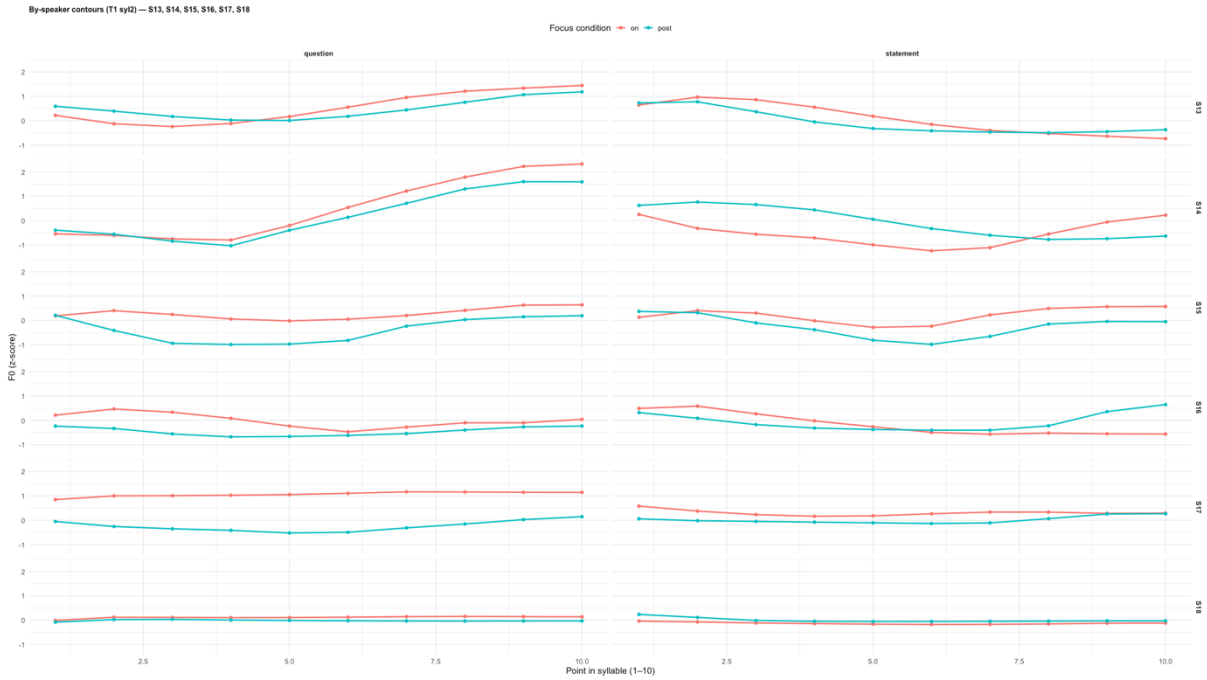
Appendix Figure 31 S.Type analysis on T1, By-speaker contours (Ch06-Ch10)



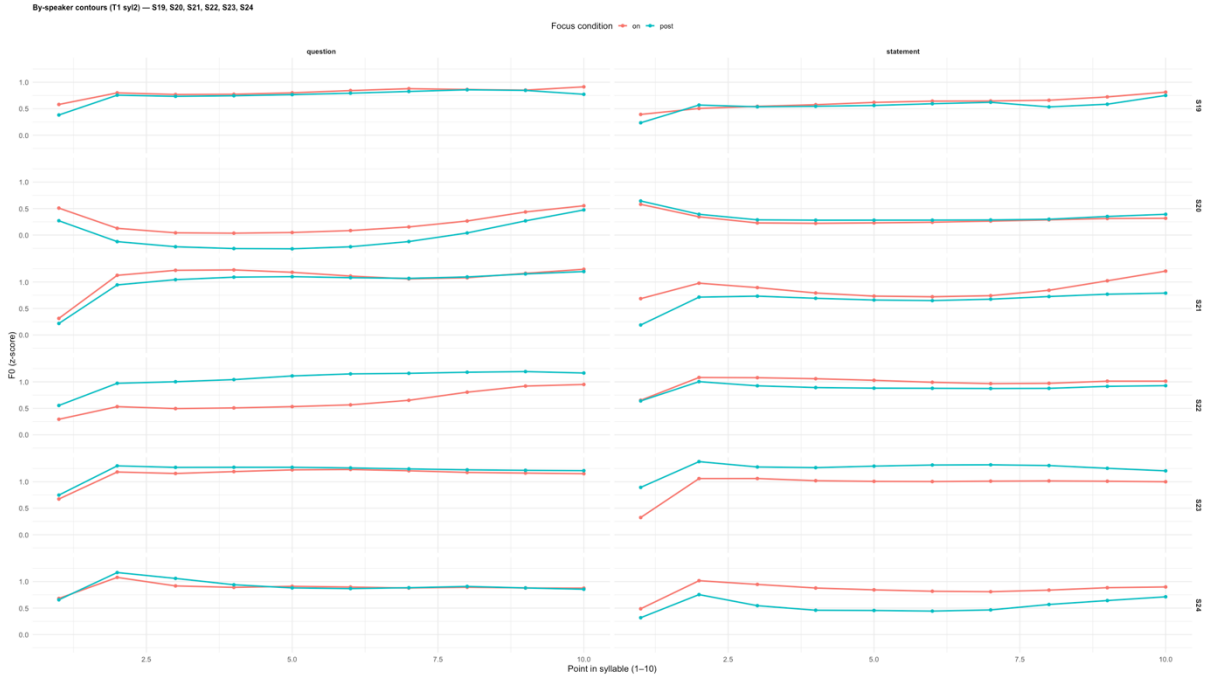
Appendix Figure 32 S.Type analysis on T1, By-speaker contours (S01-S06)



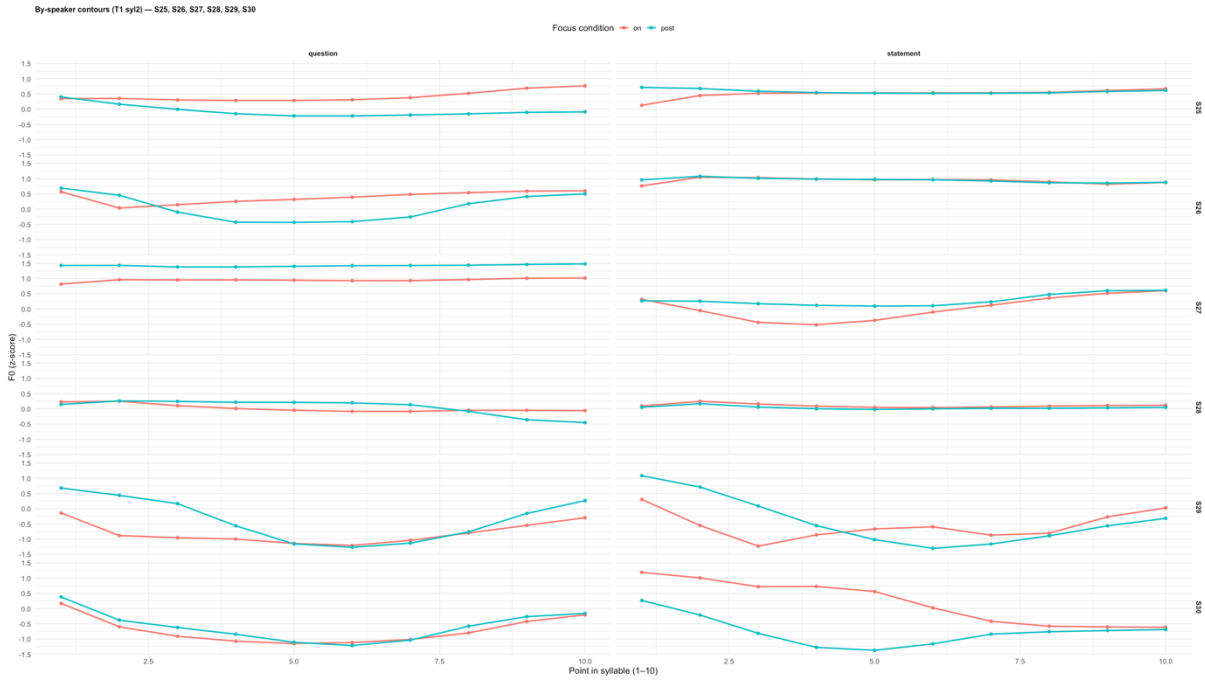
Appendix Figure 33 S.Type analysis on T1, By-speaker contours (S07-S12)



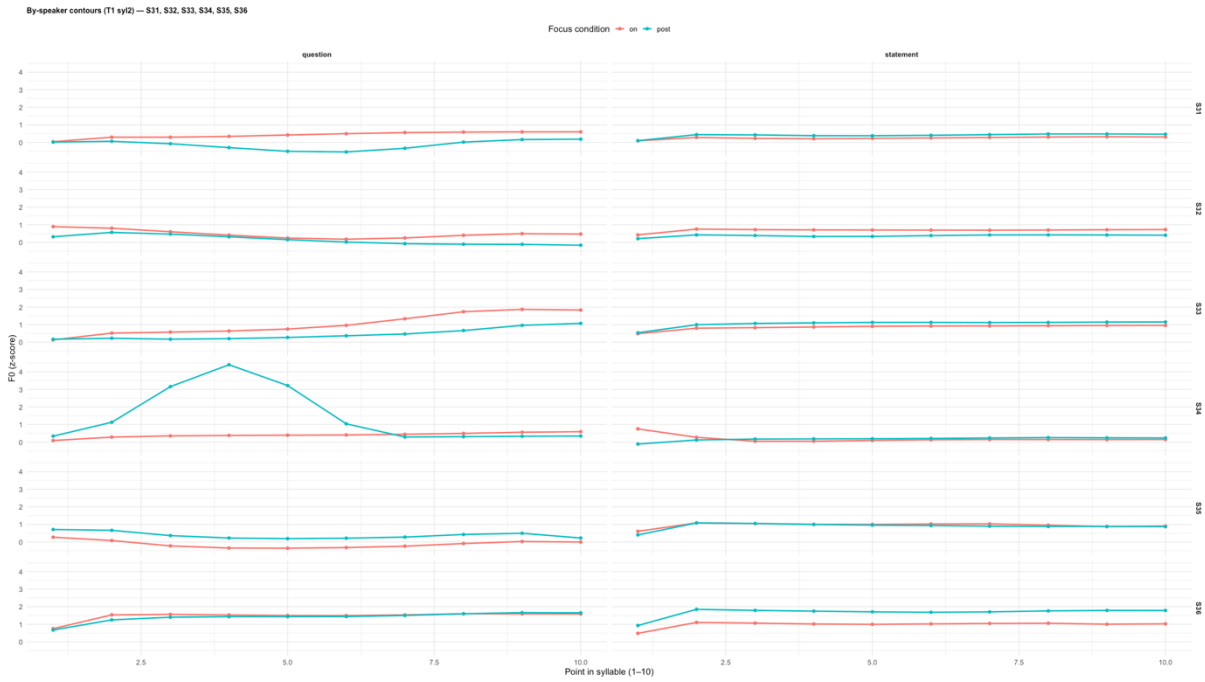
Appendix Figure 34 S.Type analysis on T1, By-speaker contours (S13-S18)



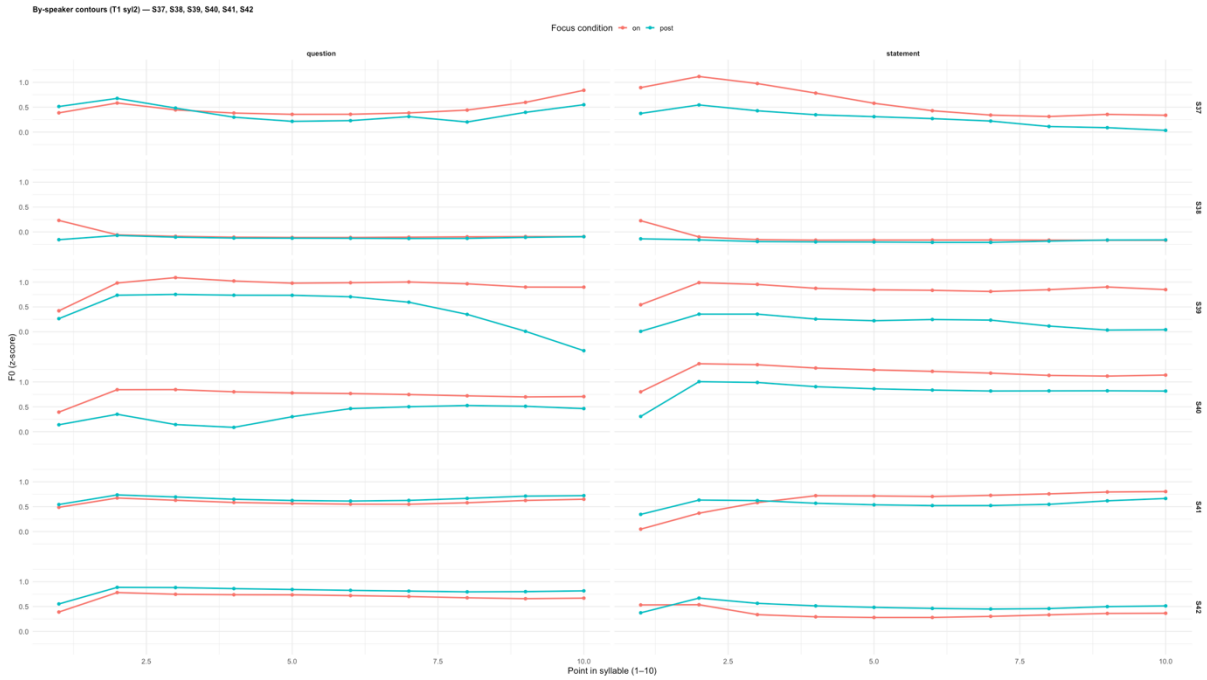
Appendix Figure 35 S.Type analysis on T1, By-speaker contours (S19-S24)



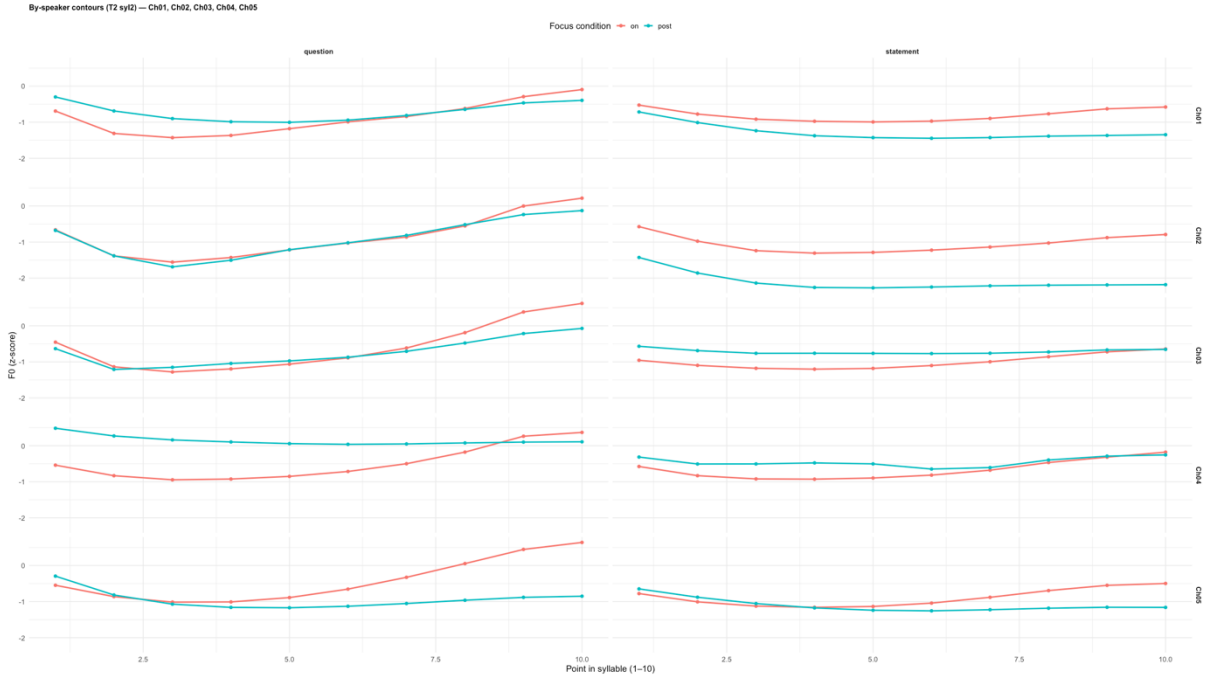
Appendix Figure 36 S.Type analysis on T1, By-speaker contours (S25-S30)



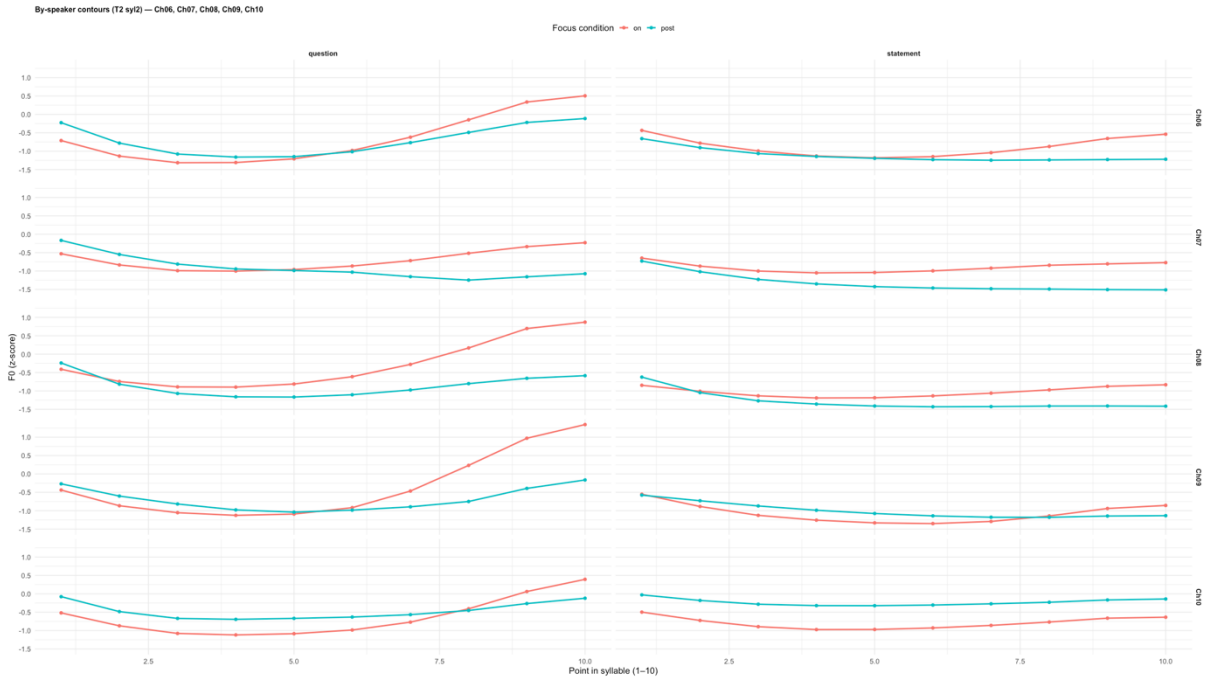
Appendix Figure 37 S.Type analysis on T1, By-speaker contours (S31-S36)



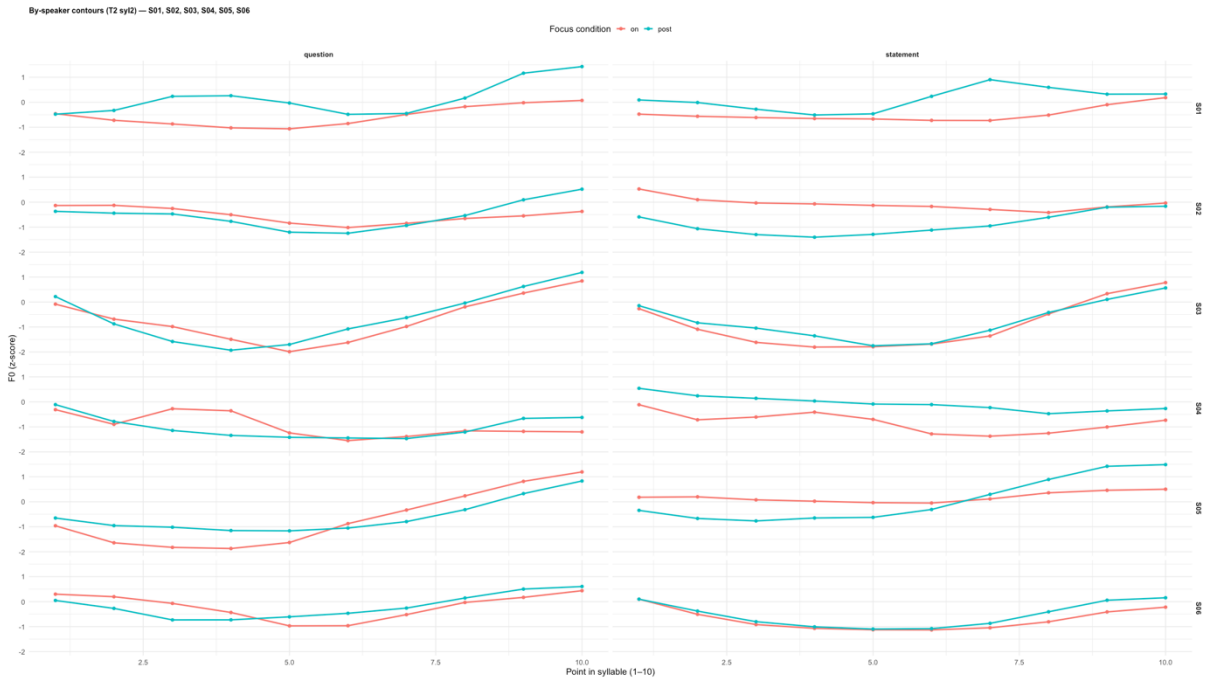
Appendix Figure 38 S.Type analysis on T1, By-speaker contours (S37-S42)



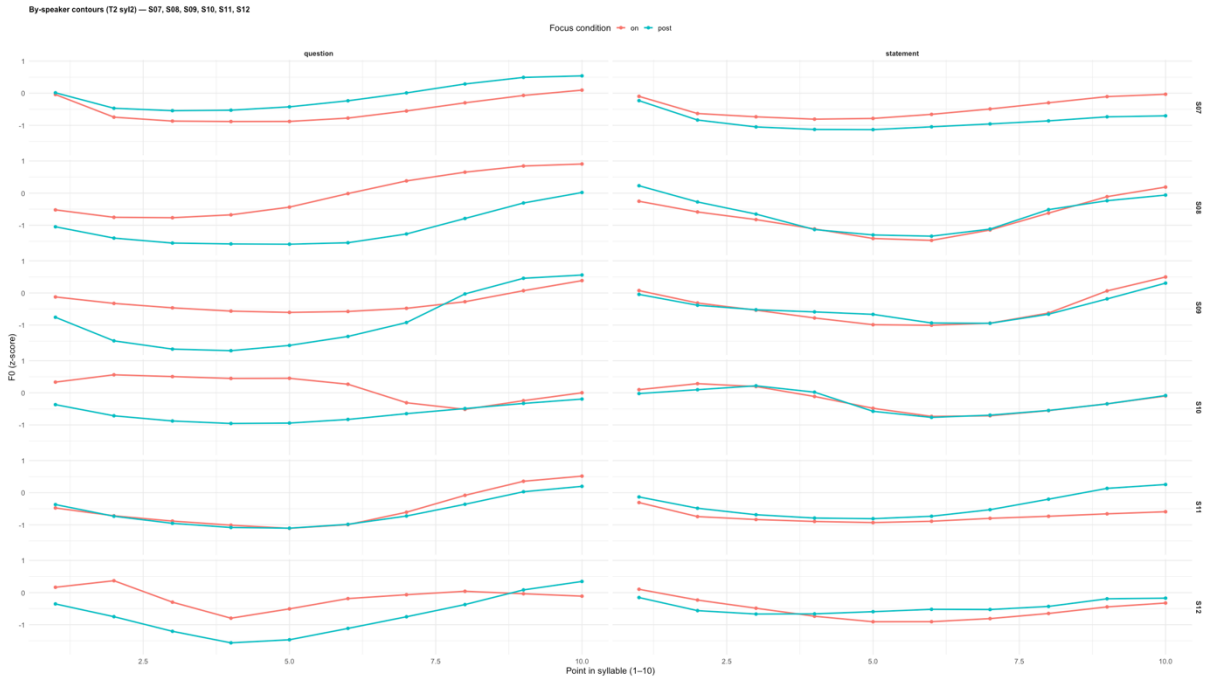
Appendix Figure 39 S.Type analysis on T2, By-speaker contours (Ch01-Ch05)



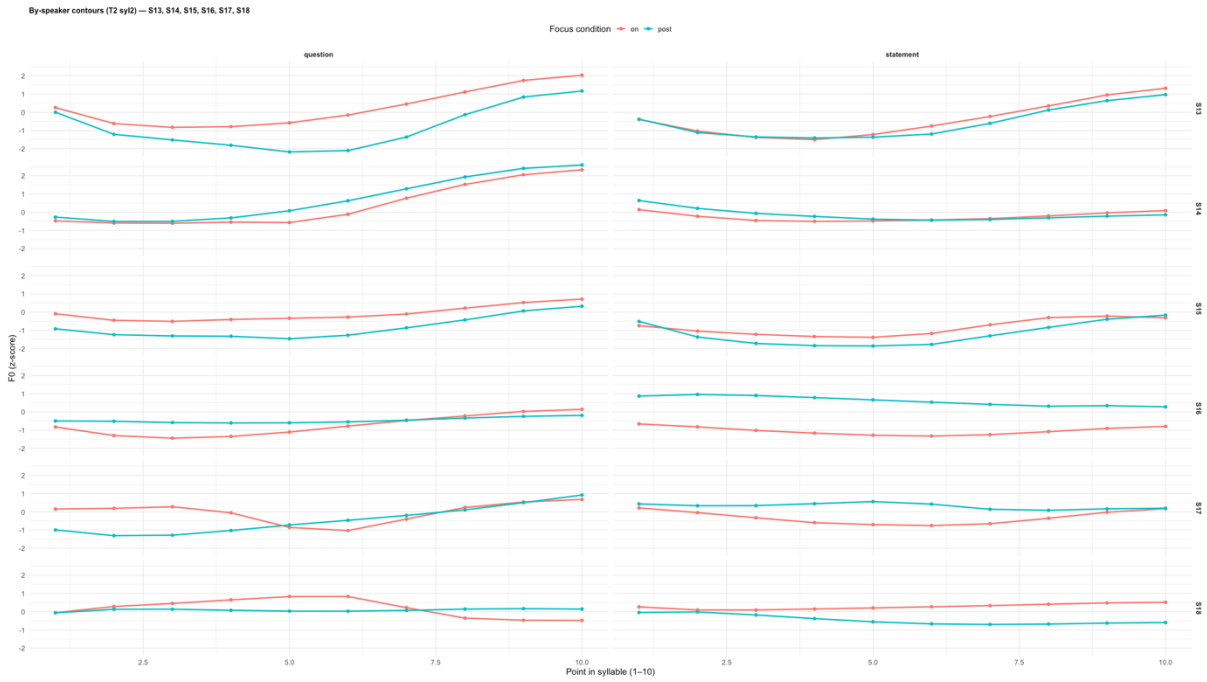
Appendix Figure 40 S.Type analysis on T2, By-speaker contours (Ch06-Ch10)



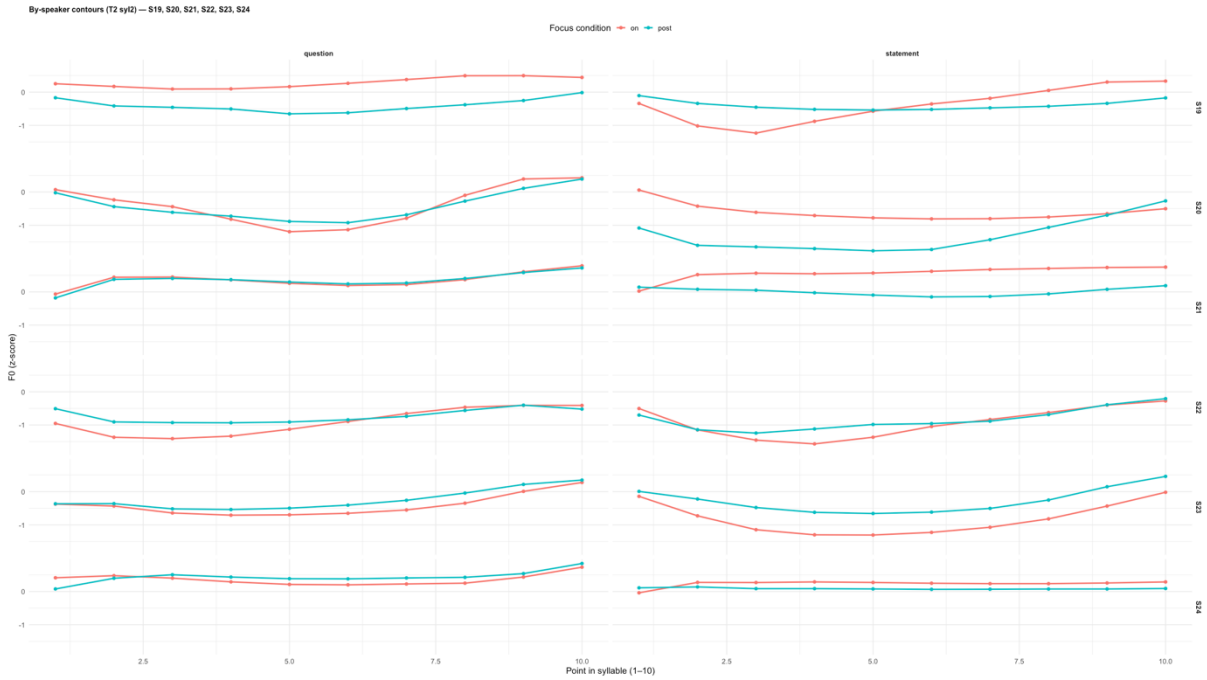
Appendix Figure 41 S.Type analysis on T2, By-speaker contours (S01-S06)



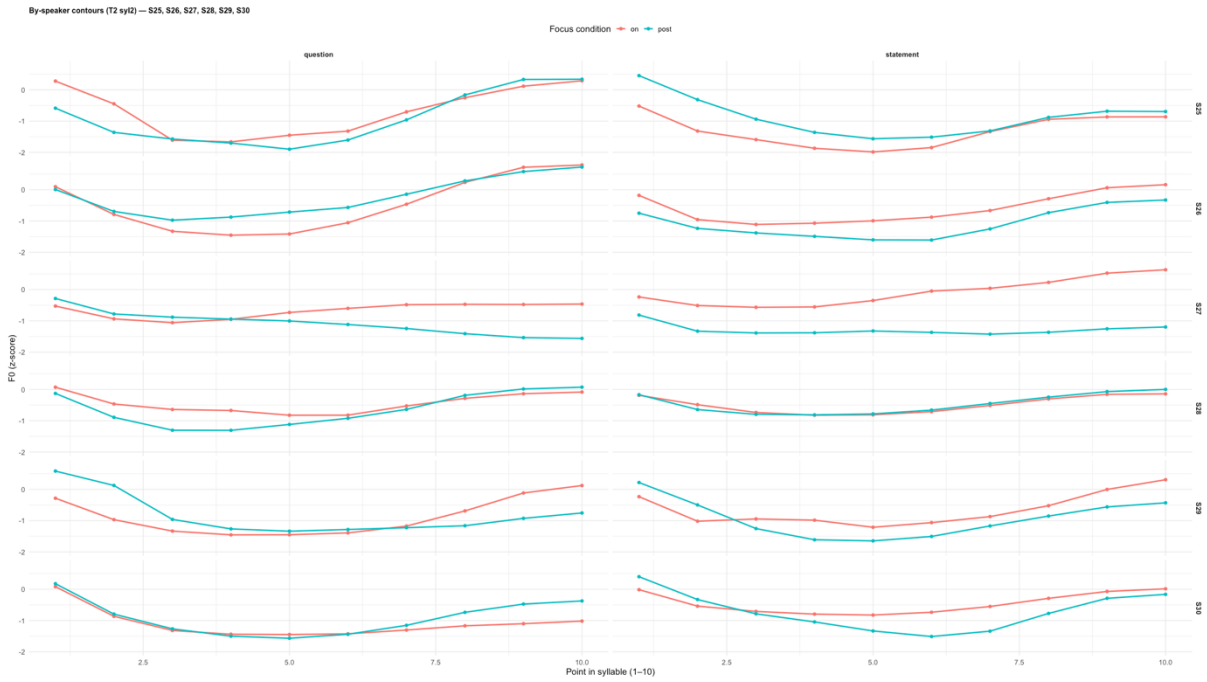
Appendix Figure 42 S.Type analysis on T2, By-speaker contours (S07-S12)



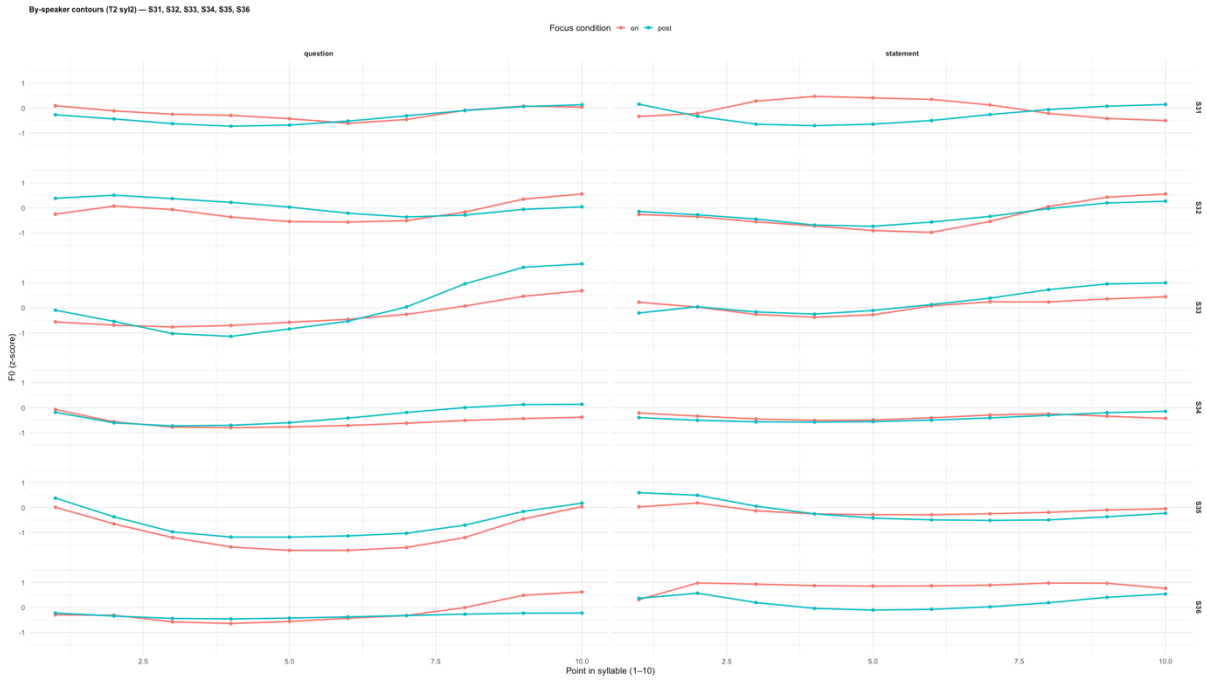
Appendix Figure 43 S.Type analysis on T2, By-speaker contours (S13-S18)



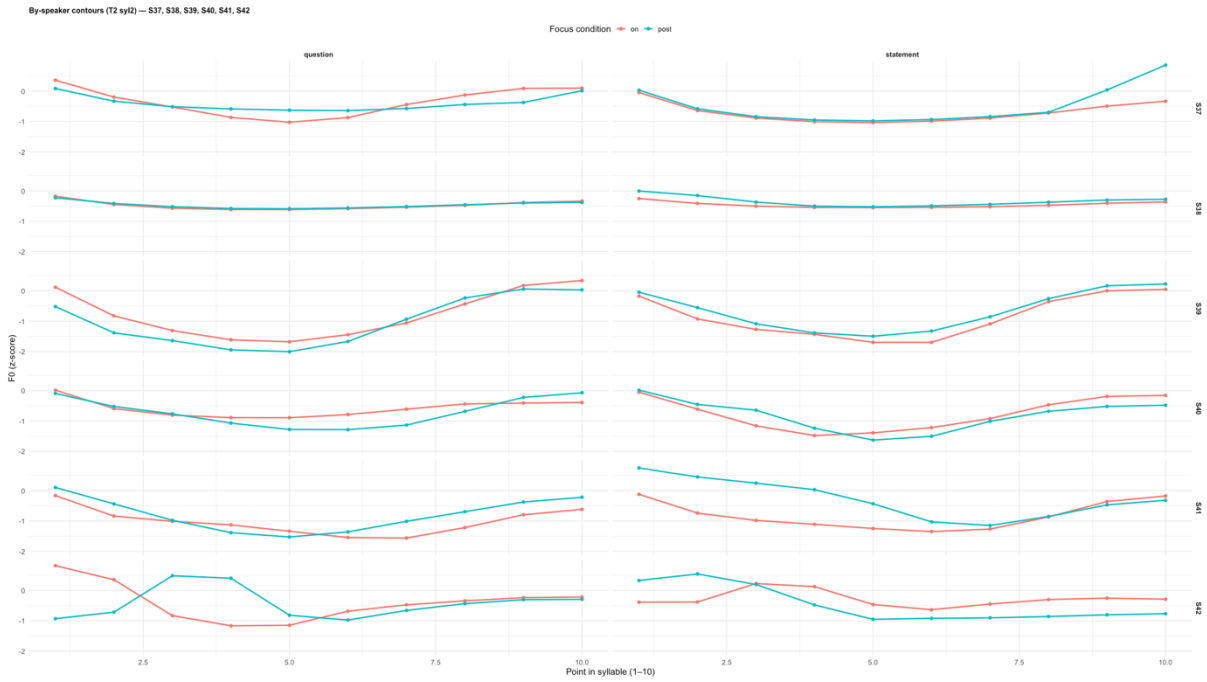
Appendix Figure 44 S.Type analysis on T2, By-speaker contours (S19-S24)



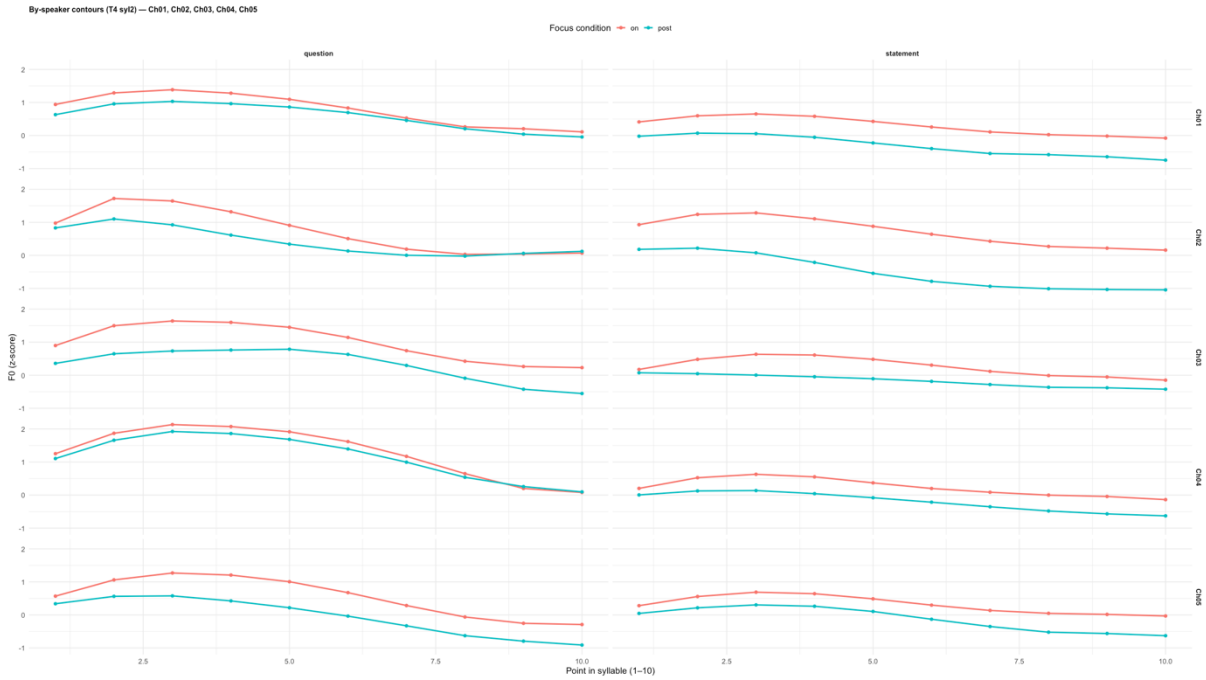
Appendix Figure 45 S.Type analysis on T2, By-speaker contours (S25-S30)



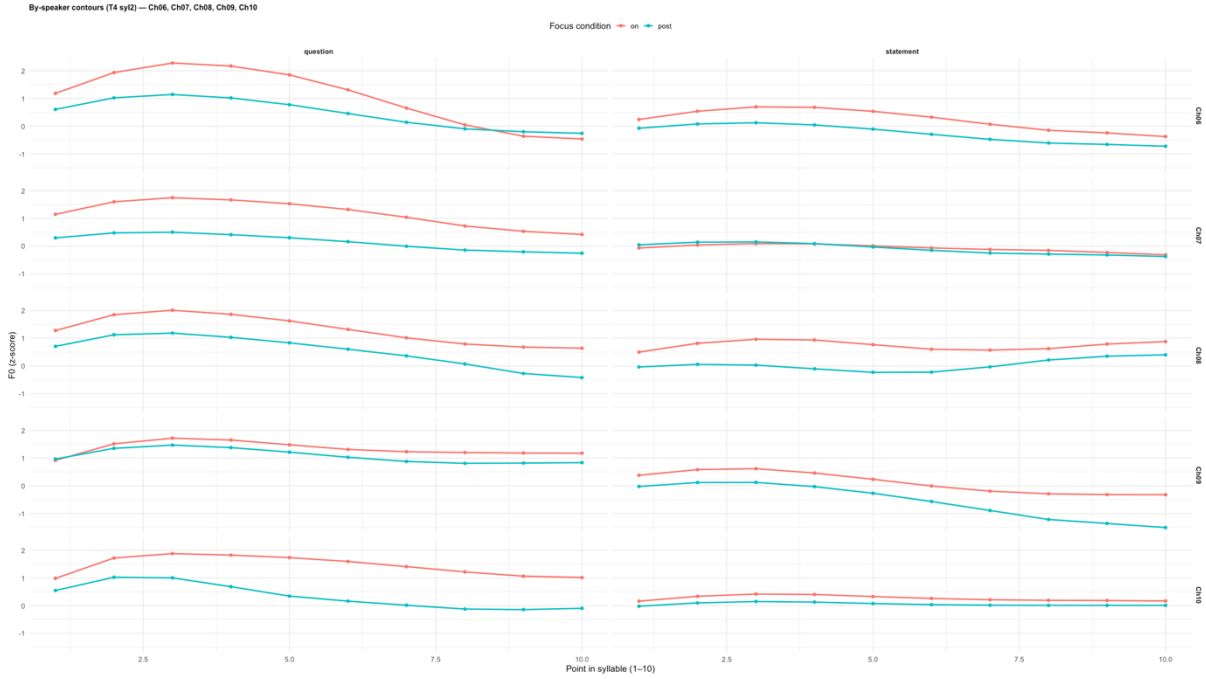
Appendix Figure 46 S.Type analysis on T2, By-speaker contours (S31-S36)



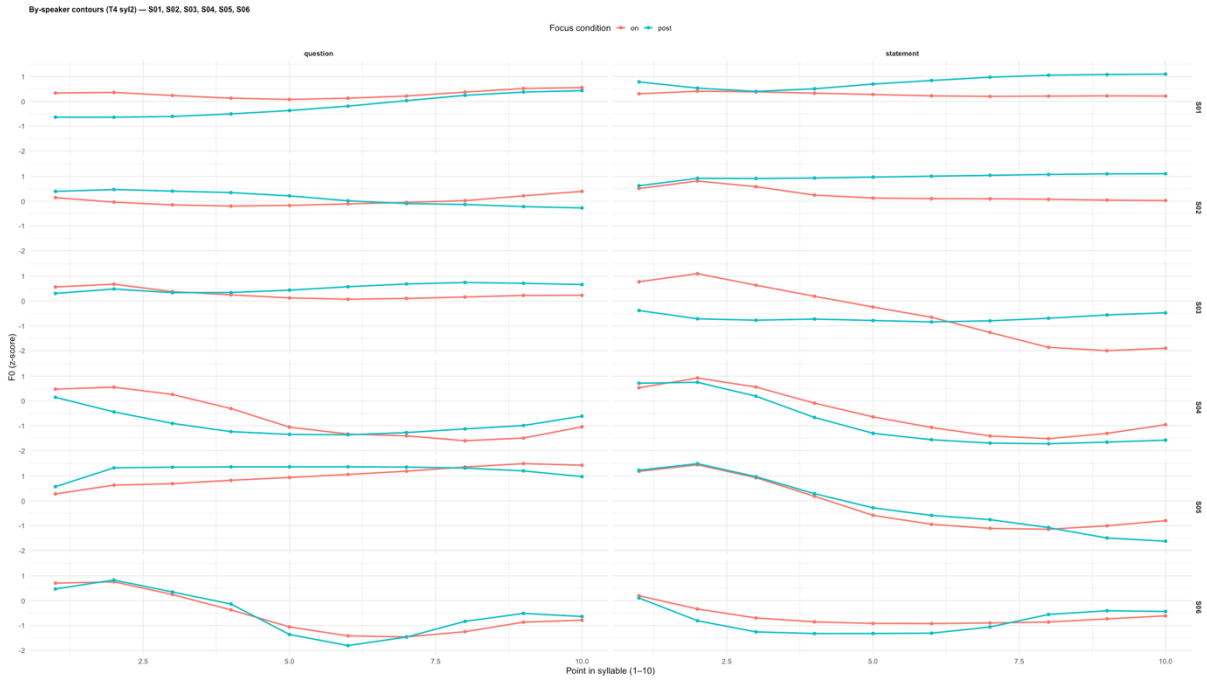
Appendix Figure 47 S.Type analysis on T2, By-speaker contours (S37-S42)



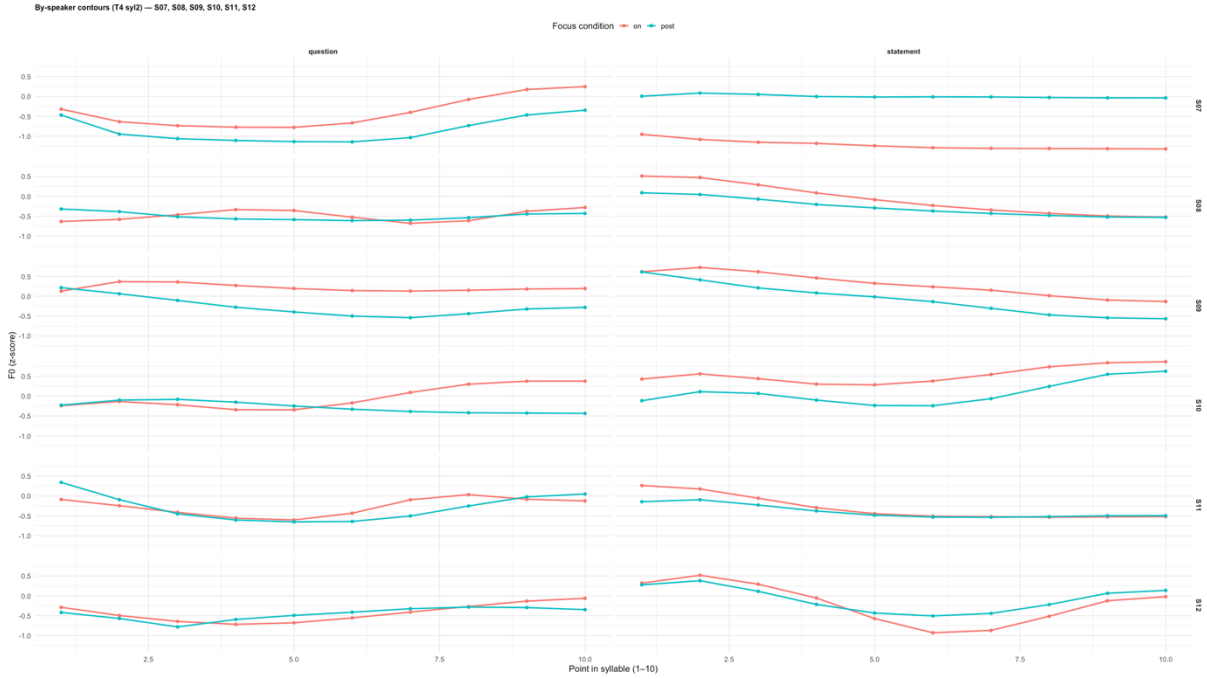
Appendix Figure 48 S.Type analysis on T4, By-speaker contours (Ch01-Ch05)



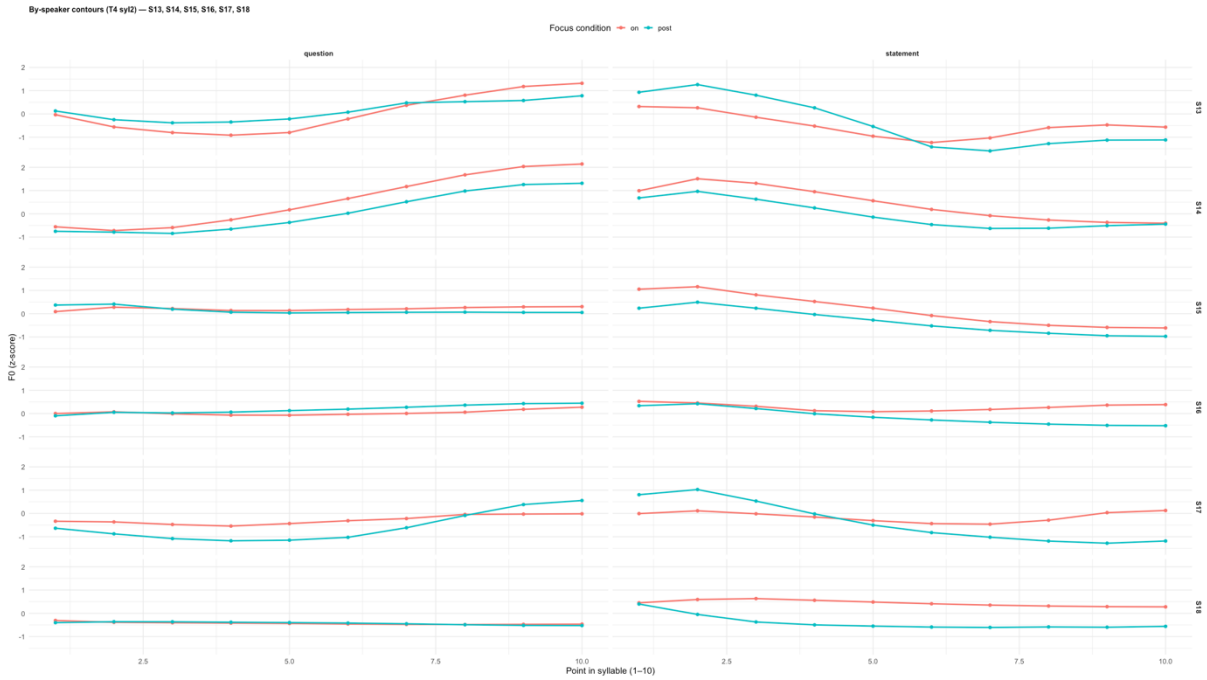
Appendix Figure 49 S.Type analysis on T4, By-speaker contours (Ch06-Ch10)



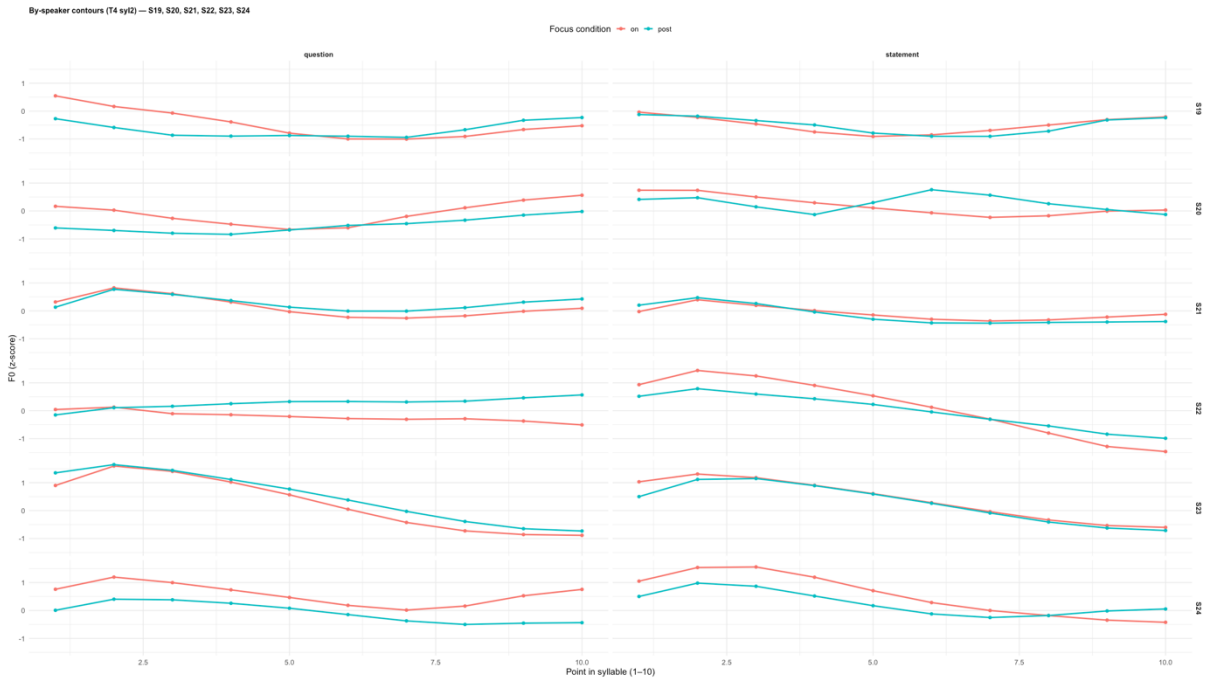
Appendix Figure 50 S.Type analysis on T4, By-speaker contours (S01-S06)



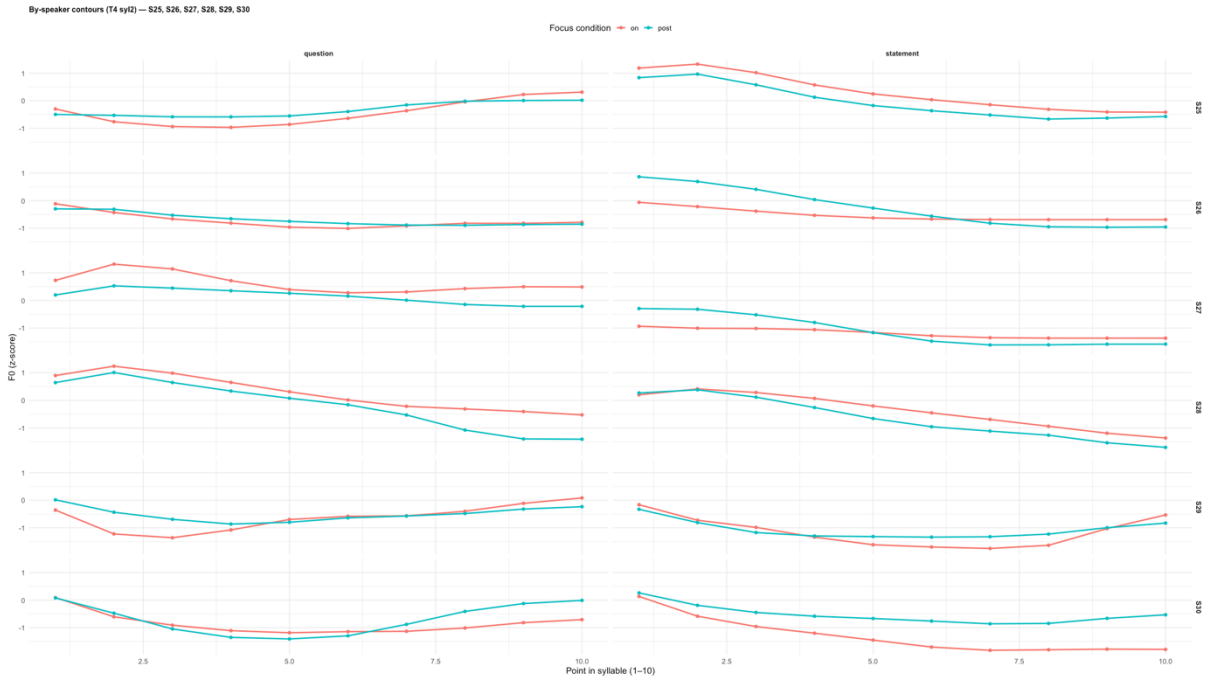
Appendix Figure 51 S.Type analysis on T4, By-speaker contours (S07-S12)



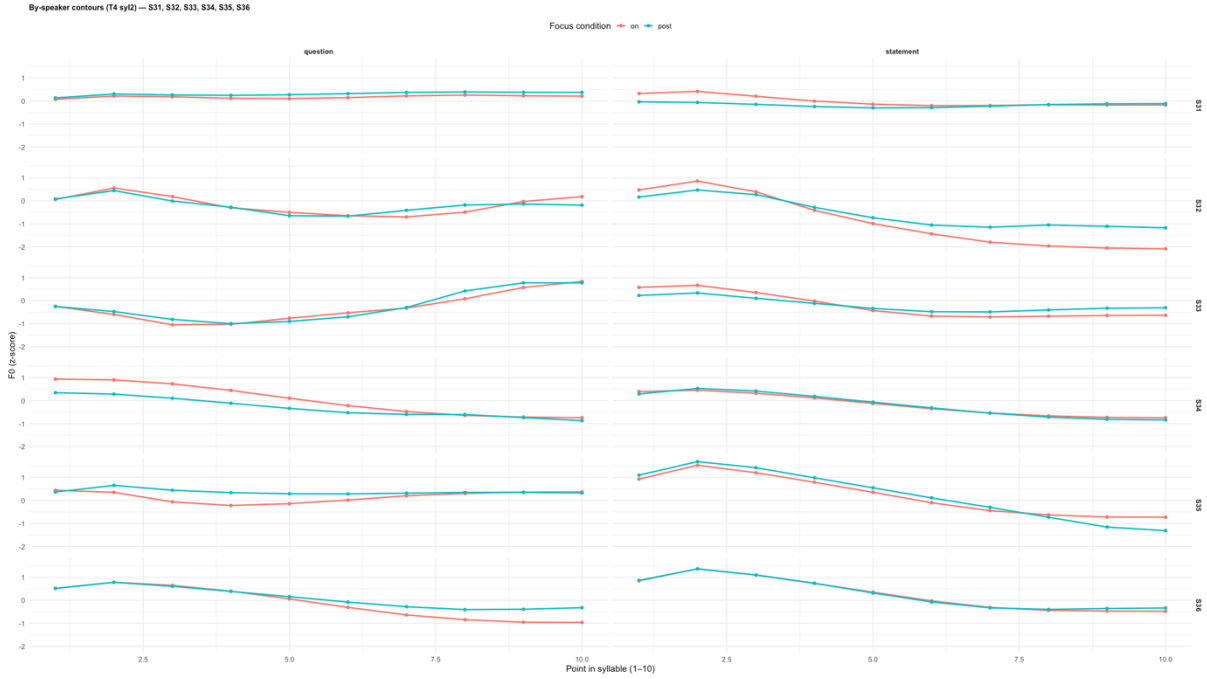
Appendix Figure 52 S.Type analysis on T4, By-speaker contours (S13-S18)



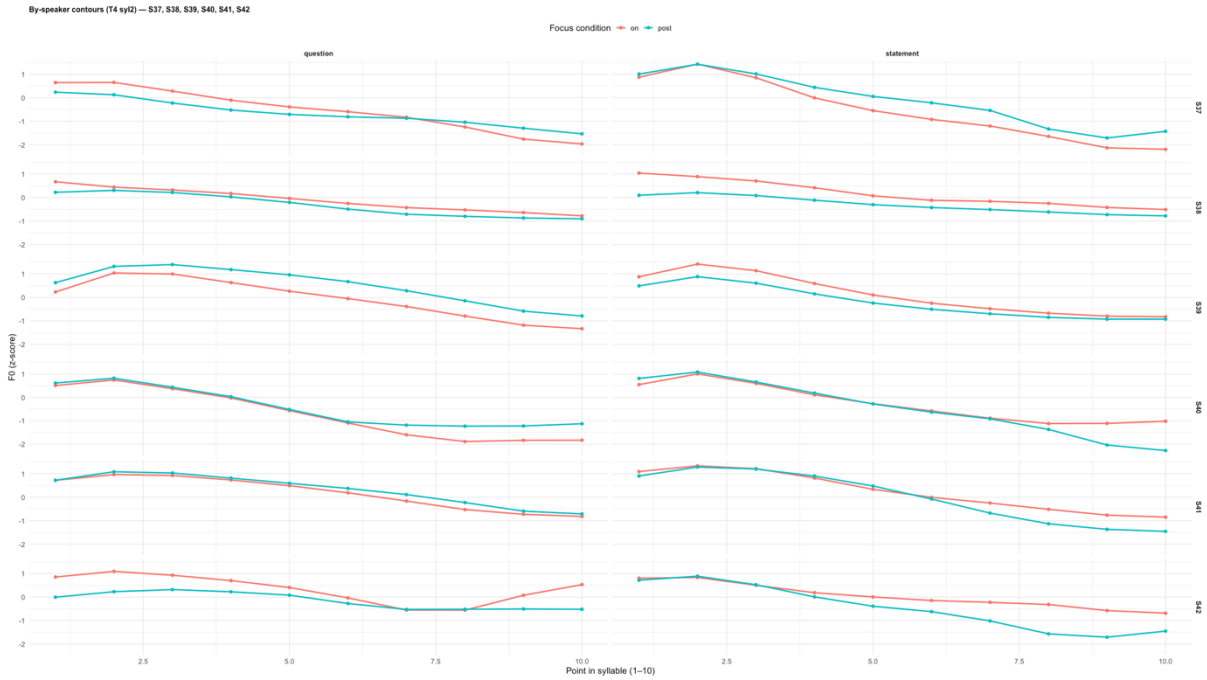
Appendix Figure 53 S.Type analysis on T4, By-speaker contours (S19-S24)



Appendix Figure 54 S.Type analysis on T4, By-speaker contours (S25-S30)



Appendix Figure 55 S.Type analysis on T4, By-speaker contours (S31-S36)



Appendix Figure 56 S.Type analysis on T4, By-speaker contours (S37-S42)