Perceptual strategies underlying second language acquisition

Perceptual strategies underlying second language acquisition

Magdalena Kachlicka

Thesis submitted for the degree of *Doctor of Philosophy*

Birkbeck, University of London

Department of Psychological Sciences

August 2023

Declaration

I, Magdalena Kachlicka, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis though appropriate citation.

Abstract

The literature suggests that listeners do not pay equal attention to all available acoustic information. Instead, when perceiving speech, they place more importance on some acoustic cues than others (Francis & Nusbaum, 2002). The patterns of weights assigned to different cues appear to change with increased linguistic experience, not only in the first language (L1; Mayo et al., 2003) but also in the second language (L2; Chandrasekaran et al., 2010). However, the role of attention and salience in cue weighting is still under discussion. This thesis presents a series of experiments investigating the relationship between dimension-selective attention, salience, and cue weighting in the context of native English speakers and Mandarin Chinese learners of English. First, I compared how prior experience (language background, musical training, and their interaction) shapes cue weighting strategies and tested whether the weighting of different cues reflects the direction of attention towards them or their salience. Compared to English speakers, Mandarin speakers showed enhanced attention to and preferential use of pitch across behavioral tasks, but no increased pitch salience measured with EEG. Effects of musical training were contingent upon participants' L1. Results also demonstrated that perceptual strategies are not consistent across tasks, suggesting they are not driven by domain-general abilities. Second, since acoustic cues play different roles across languages, learning a new language might require listeners to make greater use of L1-irrelevant dimensions. I designed a targeted training focused on redirecting listeners' attention towards an L2-relevant acoustic cue. Although the observed training effects were not long-lasting, I showed that perceptual strategies in categorizing L2 prosody could be adjusted with as little as three hours of training. This finding has the potential to inform the development of L2 learning paradigms targeting specific auditory challenges experienced by learners. Overall, this thesis provides novel insights into the long-debated role of dimension-selective attention and dimensional salience in cue weighting.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

> *"Among the gross acoustic features (…) certain ones are distinctive, recurring in recognizable and relatively constant shape in successive utterances. (…) The speaker has been trained to make sound-producing movements in such a way that the phoneme-features will be present in the soundwaves, and he has been trained to respond only to these features and ignore the rest of the gross acoustic mass that reaches his ears."*
>
> Leonard Bloomfield (*Language*, 1933, p. 79).

Abstract: Speech perception can be thought of as a perceptual categorization task in which continuous streams of sounds are divided into meaningful linguistic units such as single words, phrases, or sentences, and consonant or vowel phonemes (segmental vs suprasegmental level; Holt & Lotto, 2006). However, categorization during speech perception is not a simple process of matching the sound to the template or comparing it to the ideal exemplar of the category stored in the long-term memory (Baese-Berk, Chandrasekaran, & Roark, 2022). To understand speech, listeners need to determine the boundaries between these units by mapping continuous spectral and temporal variations to appropriate linguistic categories (Holt & Lotto, 2010). Each phonological contrast can be specified by multiple acoustic cues (Lisker, 1986) and integrating the information they convey allows for adjusting the listening strategies to various signal distortions or accents (Jasmin, Tierney, Obasih, & Holt, 2023). These cues contain the information that is needed to distinguish between different classes of events, such as phonemes, words or sentences spoken by different talkers or with different intonations. The assignment of the specific item (e.g., syllable /da/) to a discrete linguistic category (i.e., perceiving it as /da/ rather than /ta/) would therefore rely on a person's ability to utilize acoustic cues present in that signal to make the correct judgement. It seems that while adapting to a wide variety of inputs, listening conditions or tasks at hand, people use available acoustic cues differently depending on their relative importance and reliability (Holt & Lotto, 2006). It has been proposed that cue weights reflect the integration of the acoustic information and measure how much listeners attend to each of these dimensions during speech perception (Holt & Lotto, 2006; Francis, Baldwin, & Nusbaum, 2000). The weights assigned to different cues are said to depend on various experiential factors (e.g., first language [L1] background, Jasmin, Sun, & Tierney, 2021) and be susceptible to training (Lim & Holt, 2011), therefore, the strategies employed during categorization might vary considerably across people. Individual differences in cue weighting patterns have been indeed tied to second language (L2) learning (e.g., Chandrasekaran, Sampath, & Wong, 2010) and cochlear implant use (e.g., Winn & Chatterjee, 2012), suggesting that these differences and their underlying mechanisms might have significant consequences for accurate inferences about the world, effective communication, and cognition.

1.1 Insights from first language acquisition

Research suggests that we are born with broad perceptual sensitivity for speech that is not necessarily constrained to any language. Indeed, before a child's first birthday, the perceptual system begins to shift from a flexible language-general system that allows them to differentiate a wide range of sounds to a more language-specific one (e.g., Werker & Tees, 1984; Werker & Polka, 1993). It seems that as soon as within the first 6 months of life, infants begin to tune their ears to L1-specific acoustic cues (Kuhl et al., 1992) and key audiovisual correspondences (Pons, Lewkowicz, Soto-Faraco, & Sebastian-Galles, 2009).

When a listener has tuned their perceptual system to their native language, they can expertly differentiate among sounds from L1 categories. However, in doing so, they inevitably lose their ability to differentiate among non-native contrasts, as the variation between non-native categories is typically uninformative for perception in the listeners' L1. For example, Polish native speakers struggle with acquisition of English vowels because English has a considerably larger number of vowels than Polish (Schwartz, Aperlinski, Jakiel, & Malarski, 2016) and allophonic experience with dental retroflex stops by English speakers hinders their learning progress of Hindi (Best, 2003; Pruitt, Jenkins, & Strange, 2005). Throughout years of exposure to native language, our auditory system becomes tuned to salient acoustic characteristics of that language. Moreover, the developing sensitivity targets acoustic cues that carry relevant information (Holt & Lotto, 2006). The importance of these cues is guided towards effective L1 speech perception and production. For example, we sharpen our perception of cues that predict phrase boundaries (e.g., pause, final lengthening and pitch reset in Mandarin Chinese, Yang, Shen, Li, & Yang, 2014), change the words' meaning (e.g., tone contour in Cantonese and Mandarin Chinese, Francis, Ciocca, Ma, & Fenn, 2008; Hao, 2018), or emphasize the importance of that word in the sentence (e.g., greater pitch movement for emphasized phrase in English; e.g., Breen, Fedorenko, Wagner, & Gibson, 2010).

Redundancy of acoustic information

Speech is very rich in acoustic information that can be used for detecting various elements of language (e.g., short duration burst and the delay to the voicing onset for stop consonant detection, Li, Menon, & Allen, 2010; syllable lengthening for intonation and phrase boundary, Ordin, Polyanskaya, Laka, & Nespor, 2017). Some of these acoustic cues can be more informative than others. Primary cues are more reliable in predicting linguistic content and would therefore be of preferential use, whereas secondary cues would support perception with complementary information. For example, for the distinction between voiced and unvoiced consonants /zi/ and /si/voice onset time (VOT) is a primary and pitch is a secondary cue (Haggard, Ambler, & Callow, 1970; Massaro & Cohen, 1976, 1977). For linguistic focus (i.e., which word within the phrase is emphasized) in English, a change in pitch height is a primary cue and a lengthening of a word is a secondary cue (Breen et al., 2010). Research also demonstrates that the information embedded in speech signals is redundant, such that multiple acoustic cues provide information about the same category. For example, VOT or fundamental frequency (F0) amongst 14 other acoustic features can be used as cues to differentiate between voiced vs unvoiced consonants (Lisker, 1986). This redundancy in encoding is observed across language structures – at the word level in syllable stress, and at the sentence level in resolving linguistic focus (i.e., which word is emphasized in the sentence) and phrase structure (i.e., how we divide the sentence into separate phrases). But it seems that despite its availability, we do not actively utilize all this information and we rely only on a selection of those cues.

Due to changes in distributional statistics of the acoustic inputs, people tend to down-weight reliance on the dimensions which are hard to detect and provide irrelevant or unreliable information. For example, in overlapping noise, some acoustic features might be masked (Holt et al., 2018). Since the primary dimension will be deemed inaccessible, people resort to secondary dimensions. Similarly, while listening to foreign-accented speech (Idemaru & Holt, 2011; Liu & Holt, 2015), the content of it can be clearly understood even though certain characteristics are distorted and do not meet the expected pronunciation standard for a particular language. We are still able to understand speech based on other available cues. This perceptual flexibility and adjustments are possible due to redundancies in the acoustic signal – this excess of acoustic information supports comprehensibility in the face of distortion.

Congenital factors

These acoustic redundancies also support listening when some aspects of auditory processing are compromised. Individuals diagnosed with amusia (also known as tone deafness) exhibit deficits in pitch processing that present as an inability to recognize or reproduce (Peretz, 2016) and memorize pitch information (Tillmann, Schultze, & Foxton, 2009) despite normal auditory cortical responses (Norman-Haignere et al., 2016) and subcortical pitch encoding (Liu, Maggu, Lau, & Wong, 2015). Lifetime experience of unreliable pitch perception results in perceptual strategies geared towards other available cues (Jasmin, Dick, Holt, & Tierney, 2020), hinting that when faced with innate limitations, our auditory system can compensate for such subtle deficits in processing by emphasizing other aspects of speech to guarantee an overall smooth listening experience. Indeed, when asked, amusics rarely report experiencing problems with day-to-day speech perception and communication (Liu, Patel, Fourcin, & Stewart, 2010). The study also revealed that amusics rely less on pitch even if in a given task pitch differences are large enough for them to detect (Jasmin et al., 2020). It could be that the perceptual strategies employed by listeners relate not only to their perceptual abilities, but more importantly also to the perceived precision of the perceptual evidence. In fact, the reliability of perceptual dimensions appears to be linked to functional brain connectivity (Jasmin, Dick, Stewart, & Tierney, 2020). Prominent reductions in connectivity between language and pitch-related brain regions for amusics confirm that compensatory mechanisms are at play. These results indicate that the employed strategies aim not only to maximize the use of available information but also to compensate for deficiencies in detecting it.

Individual differences

Since perceptual strategies are said to arise from extensive exposure to L1 inputs, we could expect the individuals from the same language background to have their cue preferences shaped in the same way. The concept of an average listener assumes some uniformity in performance, manifesting itself in the consistent use of acoustic dimensions in speech categorization. This proves to be only partially true. In addition to the differences across speakers of various languages, previous studies provide evidence of individual differences in perceptual cue weighting patterns amongst speakers of the same L1. For example, Japanese speakers differ in their reliance on absolute vs relative durations while distinguishing singleton and geminate categories in Japanese (Idemaru, Holt, & Seltman, 2012), Azerbaijani speakers show differential weights across formant space for vowel discrimination in their native language (Mokari & Werner, 2017) and English speakers differ in the extent to which they use VOT, F0 or both cues while resolving the L1 contrast between voiced and unvoiced consonants (Kong & Edwards, 2011, 2016; Shultz, Francis, & Llanos, 2012). This means that perceptual strategies are not just mere reflections of innate auditory abilities or instantiations of the "ideal listener" template.

Recent studies also documented substantial variability in categorization gradience amongst listeners. In their study, Kong and Edwards (2011) used visual analogue scaling tasks and anticipatory eye movements analysis to measure the relative weighting of VOT and F0 cues while distinguishing voiced and voiceless stop consonants. Among the stimuli continuum from /ta/ to /da/, native English speakers were shown to vary in the degree of their use of acoustic cues and presented with different patterns of responses (Kong & Edwards, 2011; Ou, Yu, & Xiang, 2021). In addition to individual differences in using primary vs secondary cues, some people seemed to have more gradient-like responses (i.e., choosing responses across the entire scale), while others displayed more categorical patterns (i.e., choosing responses mostly at the two endpoints of the scale). These differences were linked to listeners' sensitivity to the secondary cue (here F0), where increased sensitivity to that secondary cue was indicative of more gradience in their perception (Kong & Edwards, 2016; Kapnoula et al., 2017), indicating less clear boundaries between linguistic classes. This pattern was further confirmed in another study, where participants showed considerable differences in perceptual categorization responses (i.e., using VOT, F0 or both cues), even in the absence of evident differences in production (Schertz, Cho, Lotto, & Warner, 2015). Neural sources of such individual variability in the categoricity of speech perception were also identified (Fuhrmeister & Myers, 2021). In response to a fricative continuum between /sign/ and /shine/, participants with more categorical perception showed an increased surface area of the right middle frontal gyrus. Additionally, more gyrification was shown in participants with more consistent responses suggesting some link between brain structure and language proficiency. These results show that categorization during speech perception can be achieved by mapping various acoustic cues to linguistic categories in any given language, but the strategies used to perform that mapping significantly differ across listeners.

People can also extract and use phonetic information across different dimensions (Hazan & Rosen, 1991; Clayards, 2018). Such a strategy is not necessarily tied to being an expert in one particular dimension but might indicate more complex interactions in how people use different acoustic cues. In fact, a study in the visual domain showed that some participants master only one dimension necessary for accurate performance, whereas others learn two dimensions that are not equally relevant for completing the task (Shamloo & Helie, 2020). This is reflected not only in their performance but also in selected strategies because even in a subset of people who learn both dimensions, we observe some individual differences. While some participants successfully integrated across dimensions, others showed errors in their behavior, suggesting that they could use only one of the acquired cues effectively. This finding provides further evidence that not all listeners might comply with the same "ideal listener" profile. Some shared patterns are present, but individual differences also play a role. Taken together, these results suggest that the emerging perceptual profiles are not some fleeting properties but are indeed stable in time and might represent a consistent pattern of each listener's speech perception. What drives these differences remains an open question.

1.2 Perceptual strategies in second language learning

Mapping acoustic variation onto various linguistic categories is already complex in our first language but becomes even more difficult when learning a second language because acoustic cues can play different roles across languages. For example, Japanese native speakers find it difficult to distinguish between the /u/, /y/, and /ø/ vowels in French since there are no such categories in their L1 phonological inventory and they lack sensitivity to the third formant (F3), which is generally not useful for distinguishing phonemes in Japanese but is crucial for resolving vowel contrasts in French (Kamiyama & Vaissiere, 2009; Kamiyama, 2011). In the same way, native English speakers have not learned to distinguish between various ways of articulating lexical tones marked by changing pitch contours, which, on the other hand, is crucial for conveying meaning in tonal languages (Francis, Ciocca, Ma, & Fenn, 2008; Hao, 2018). As a result, an optimal L1 listening strategy might not be as effective for any subsequent L2.

Perceptual strategies are language-specific

The literature suggests that listeners do not pay equal attention to all acoustic information available to them. Instead, when perceiving speech, listeners seem to place more importance on some acoustic cues than others (Francis, Kaganovich, & Driscoll-Huber, 2008). Furthermore, the patterns of weights to different acoustic cues appear to change with increased linguistic experience, not only in L1 (e.g., Mayo, Scobbie, Hewlett, & Waters, 2003) but also in L2 (e.g., Chandrasekaran et al., 2010). However, the role of attention in cue weighting is still under discussion and little evidence was provided in support of this hypothesis. As discussed in the previous section, individuals' ability to utilize various cues is presumed to change as a function of extensive exposure to native language inputs. That suggests that listeners shift their ear's susceptibility to learning perceptually essential contrasts in their L1 while neglecting dimensions that might be useful for acquiring contrasts in subsequent languages. Such narrowed specialization within the L1 phonological system comes at a price of language-specific differences in using various acoustic cues (e.g., differences between German, Italian and Spanish monolinguals, and Spanish-Basque bilinguals, Ordin et al., 2017), neural encoding of sound (e.g., Mandarin late learners of English vs American English native speakers, Krishnan, Xu, Gandour, & Cariani, 2005) or speech perception more generally (e.g., Finish vs Estonian monolinguals, Naatanen et al., 1997).

Phonetic categorization

The differences in how individuals use acoustic cues manifest themselves during speech perception at the segmental level (i.e., during phonetic categorization, e.g., consonants' voicing continuum, McQueen, 1996, Francis, Kaganovich, & Driscoll-Huber, 2008; vowel continuum, Kivisto-de Souza, Carlet, Julkowska, & Rato, 2017). For example, cues such as VOT, F0 at vowel onset and closure durations are used by speakers of Korean and English to detect stop contrasts in their native languages, but to a different extent. While in Korean equal weights are given to F0 and closure duration (e.g., Silva, 2006; Lee & Jongman, 2019), VOT is the primary cue in English (Schertz et al., 2015). Compared to Spanish speakers, English natives rely more on first formant (F1) onset frequency (Schertz, Carbonelli, & Lotto, 2020). Furthermore, while categorizing Dutch vowel contrasts, Dutch, German and Spanish speakers differentially weigh vowel spectrum and duration (Escudero, Benders, & Lipski, 2009). Both Dutch and German native speakers relied more on spectral features than duration, whereas Spanish speakers favored duration. Authors suggested that these differences can be explained in terms of the cross-linguistic comparison between German and Dutch vowels – spectral differences are indeed more important for distinguishing vowel contrasts in Dutch and German. The relative size of the vowel inventory was also shown to relate to speakers' reliance on durational cues (i.e., the larger the vowel space, the stronger the reliance on duration cues; Kivisto-de Souza et al., 2017).

Prosodic categorization

Consistent L1-driven differences were also shown cues at the suprasegmental level (i.e., during prosodic categorization, e.g., phrase boundaries, Kuang, Chan & Rhee, 2022; intentions and affect, Hellbernd & Sammler, 2016). Ordin and colleagues (2017) demonstrated that German and Spanish-Basque speakers' speech segmentation in an artificial language was improved by word-final lengthening, whereas Italians benefited most from penultimate syllable lengthening. Furthermore, antepenultimate lengthening impeded boundary detection by Spanish and Italian speakers, showing that the interpretation of syllable lengthening in detecting a final phrase boundary is not universal (Ordin et al., 2017). Amongst the three cues for intonational phrase boundary in Chinese (pause, final lengthening, and pitch contour reset), pause was shown to be more strongly weighted than the other two cues (Yang et al., 2014). Although pitch reset and final lengthening were perceptually equivalent, the effects of these cues were not cumulative. Acoustic cues used for stress identification also vary across languages (for review see: Gordon & Roettger, 2017). For instance, English and Mandarin native speakers used vowel quality as the primary and pitch as a secondary cue, while pitch was completely disregarded by Russian speakers when identifying stress location in a disyllabic nonword "maba" that was phonologically and phonotactically permissible across these languages (Chrabaszcz, Winn, Lin, & Idsardi, 2014). Russian speakers relied more on duration and intensity cues instead. Besides coarticulation (e.g., Repp, 1983) and intonation effects (e.g., Peng, Lu, & Chatterjee, 2009), systematic interactions between cues for individual phonemes and speech prosody are also present (e.g., de Pijper & Sanderman, 1994; Reinisch, Jesse, & McQueen, 2011).

Tonal vs non-tonal languages

One striking difference in cue use can be observed between tonal and non-tonal language speakers. In tonal languages, contrastive fundamental frequency variation is used to mark lexical tones; this conveyance of meaning diverges from functions of pitch in English or any other non-tonal language (Francis et al., 2008; Hao, 2018). In Mandarin Chinese, each syllable has one of four different pitch contours changing the meaning of the word (e.g., van de Weijer & Sloos, 2014). Depending on how the speaker pronounces syllable /ma/ it could mean either "mother" (first tone – T1, high and flat pitch contour), "hemp" (second tone – T2, rising pitch), "horse" (third tone – T3, falling-rising or dipping pitch) or "scold" (fourth tone – T4, falling pitch). Pitch contour is therefore a vital cue to the meaning of words in tonal languages, whereas in non-tonal languages it serves a more secondary role in signalling linguistic focus (greater pitch movement for emphasized words; e.g., Breen et al., 2010), emotional state (various patterns of pitch levels and contour types linked to different emotions; e.g., Rodero, 2011), speakers' intentions (e.g., exaggerated pitch patterns for sarcasm and irony; Attardo, Eisterhold, Hay, & Poggi, 2003; increased pitch during attempted deception; Streeter et al., 1977) or sentence type (e.g., rising pitch contour for questions; Bartels, 1999). Some linguists consider tones at the segmental level along with consonants and vowels[1] (e.g., Duanmu, 1994), although their influence clearly extends to the suprasegmental level (So & Best, 2010; Li, Ong, Tuninetti, & Escudero, 2018; for a broader discussion on this topic see Best, 2019) as they not only change the word meaning but inevitably also sentence prosody, affecting the overall melody of a language.

Therefore, a question arises whether such extensive experience with pitch variation can change how tonal language speakers process sound. Research provides abundant evidence supporting this hypothesis showing tonal language speakers' expertise in pitch processing. Mandarin native speakers were shown to be more precise in pitch discrimination (e.g., Giuliano et al., 2011; Liu, Hilton, Bergelson, & Mehr, 2023) and pitch interval discrimination (quantified as the relative pitch difference between two consecutive pitches;

[1] Tones are mostly manifested on vowels, but their influence extends to voiced consonants.

e.g., Pfordresher & Brown, 2009; Hove, Sutherland, & Krumhansl, 2010; Giuliano et al., 2011; Creel et al., 2018; Zheng & Samuel, 2018), to be better at melody discrimination (Bidelman et al., 2013; Wong et al., 2012), and to have more sensitivity to musical pitch contour (Bradley, 2012). Prior experience with a tonal language was also shown to facilitate the learning of subsequent tonal languages (Qin, Zhang, & Wang, 2021). Differences were observed not only in speech perception, but in production as well, showing comparable use of amplitude and duration with higher overall F0 (Zhang, Nissen, & Francis 2008). Keating and Kuo (2012) steered away from interpreting Mandarin speakers' use of a larger F0 range and higher F0 peak because their language has multiple tones. Instead, they justified these differences by how Mandarin speakers pronounce their high-falling tone, i.e., increased range with a high F0 peak that enhances the falling pitch contour (Keating & Kuo, 2012).

However, some researchers claimed that, contrary to expectations, tone language fluency has detrimental effects on pitch discrimination abilities (e.g., Stagray & Downs, 1993; Bent, Bradlow, & Wright, 2006; Peretz, Nguyen, & Cummings, 2011). Tonal language speakers were shown to have trouble recognizing downward pitch changes for music pitch intervals ranging from 1 to 15 Hz (Peretz et al., 2011) and identifying rising, falling or flat non-speech tones (Bent et al., 2006). However, it is worth noting that both studies used stimuli from outside the usual Mandarin speakers' repertoire for testing (i.e., musical notes or non-word stimuli) and they performed significantly better than English native speakers in a speech tone discrimination condition (Bent et al., 2006). In another study on musical pitch perception (Pfordresher & Brown, 2009), Mandarin speakers were better at imitating musical pitch and discriminating intervals, but the effect of single pitches was limited to production only (no effects in musical notes discrimination condition).

The presence of such conflicting findings can be a consequence of different pitch aspects being investigated across various types of stimuli (i.e., not all the aspects of pitch processing might matter equally to tonal language speakers, and their relevance might differ across contexts). Researchers recognized that the selection of task and pitch context might in fact be responsible for yielding discrepant results (Keating & Kuo, 2012)[2]. The relative difference between pitches, pitch movements or changing pitch contours that signal meaningful information in L1 might be more relevant than pitch height. Indeed, in nonsense words Mandarin speakers give more weight to pitch movements signaling lexical contrasts than Dutch speakers, but do not show increased weights for post-lexical contrasts signaling no meaningful information (Braun & Johnson, 2011). These results are also consistent with the idea that pitch height and contour are evaluated separately (Massaro, Cohen, & Tseng, 1985). Research also hinted that the benefits of tonal language might be limited to low-level pitch perception (Bidelman, Gandour, & Krishnan, 2011a). Musical notes or nonsense stimuli might also be detected with strategies that are not necessarily mirrored from speech perception. It is possible that the pitch advantage of tonal language speakers applies to meaningful lexical contrasts but does not extend to other stimuli where other cues might be more relevant (e.g., Bradley, 2012; but see Jasmin et al., 2020 for cue-weighting patterns in musical beat perception showing over-weighting of pitch when contrasted with duration cues). However, research provides evidence that the effects of musical training and tonal language experience on pitch processing might be bidirectional (e.g., Bidelman, Hutka, & Moreno, 2013; Tang et al., 2016; Qin, Zhang, & Wang, 2021).

Differences in cue weighting

Studies in cue weighting test the relative importance placed on information gained from pitch versus from other dimensions. A study comparing the importance of pitch vs duration demonstrated that Mandarin Chinese speakers not only weigh pitch more compared to native English and Spanish speakers, but also struggle to ignore pitch information during speech and non-speech categorization tasks, even when explicitly asked to do so (phrase boundary and linguistic focus categorization tasks; Jasmin, Sun, & Tierney, 2021) and

[2] It is worth noting that research comparing tonal vs non-tonal languages very often combines speakers of various languages within each of those groups. These languages and their phonological systems might differ from one another quite substantially contributing to growing inconsistency in the observed results (a detailed discussion of these differences is beyond the scope of this thesis but see Best, 2019 for a short overview of pitch roles across various tonal and non-tonal languages).

overuse pitch when speaking (Nguyen, Ingram, & Pensalfini, 2008; Zhang, Nissen, & Francis, 2008). While distinguishing English statements vs questions, Chinese listeners were not affected by amplitude and duration changes as English speakers were, suggesting differential reliance on these cues (Feng et al., 2019). Expertise with pitch contour also influences lexical stress perception. In an event-related potential (ERP) study investigating the perception of syllable stress and use of pitch, duration and intensity in the disyllabic non-word /dede/ English and Mandarin native speakers demonstrated different response patterns to iambic (unstressed syllable followed by stressed syllable) vs trochaic stress placement (stressed syllable followed by unstressed syllable; Wang, 2008). English speakers were similarly sensitive to pitch changes located at the initial and final syllables of the word and more sensitive to the duration of the final syllable. Mandarin speakers on the other hand were similarly sensitive to the duration at both positions and more sensitive to pitch at the final position, showing a somewhat reverse pattern to English speakers. At the same time, both groups were more sensitive to intensity change at the second syllable, suggesting differential perception of duration and pitch cues only. Further research confirmed this pitch specialization in Mandarin speakers by showing that English speakers use pitch, duration and intensity in stress perception (disyllabic non-words, Zeng et al., 2020; real words, pseudowords and hums, Yu & Andruski, 2010), whereas for Mandarin speakers pitch was a decisive cue. These results indicate a transfer of F0 weighting to L2 rather than just using tones at the phonological level.

More insight into L2 English stress perception was offered by a study manipulating the presence of acoustic cue pairs (vowel quality vs intensity, vowel intensity vs duration) in natural vs flat pitch conditions (Zhang & Francis, 2010). It seems that when vowel quality is available as a cue, Mandarin speakers rely on it to a similar extent as English speakers, even though it is not a decisive cue for detecting lexical differences in their native language. Mandarin speakers also ignore intensity cues when vowel quality is available. It is possible that Mandarin listeners use vowel quality as a cue to English lexical stress instead of perceiving it in terms of a prosodic contrast. This interpretation conflicts with the proposition that cues featured in L2 but not in L1 phonological system will remain unattended or down-weighted (e.g., Francis & Nusbaum, 2002; Iverson et al., 2003; Guion

& Pederson, 2007). An alternative explanation suggests that since Mandarin speakers do not use vowel quality to resolve lexical contrasts in their native language, they cannot use their default native strategies in L2 English. That forces them to use cues they can easily detect to resolve the perceptual problem at hand (here: vowel quality), suggesting that even though indirectly, our L1 strategies might shape the way we listen to sounds in subsequent languages.

Differences in neural pitch encoding

Effects of speaking a tone language on verbal and non-verbal pitch perception and neural encoding are well documented. Neural imaging data support the view that different aspects of pitch perception might be processed separately by showing lateralization effects and hemispheric specialization for contour and interval cues for pitch (Trainor, McDonald, & Alain, 2002). Mandarin native speakers displayed stronger responses to pitch contours in the right hemisphere and to pitch intervals in the left hemisphere, whereas no such specialization was observed for native English speakers (Bidelman & Chung, 2015). High-density electrocorticography (ECoG) recordings from the exposed temporal lobe during surgery showed enhanced neural processing of Mandarin tone categories (Li et al., 2021). This study considered two aspects of pitch relevant in tone perception, namely speaker-normalized relative pitch height and pitch change and demonstrated that they might be encoded independently. To perceive lexical tones, listeners need to integrate information from pitches that are variable across speakers and also intonational information that varies across utterances. Results revealed that speaker-normalized pitch is not language-specific, as it can be used, to a smaller degree, by English speakers when processing sentence prosody and intonation. On the other hand, Mandarin speakers showed different temporal response function (TRF) tuning curves for pitch features reflecting the differences between the two languages (i.e., wider pitch range and lower relative pitch).

Distinctive patterns of responses to various aspects of pitch were also seen when comparing the performance of tonal language speakers to non-tonal musicians hinting at the relationship between musical and language experience and its role in pitch processing. Tone language speakers and musicians have a superior sensory encoding of pitch-relevant information at both subcortical (Bidelman, Gandour, & Krishnan, 2011b) and cortical

(Chandrasekaran, Krishnan, & Gandour, 2009) levels, but the differential benefits emerge when considering more complex aspects of pitch and music perception. A mismatch negativity (MMN) study compared tonal native speakers to non-tonal musicians on detecting deviants in pitch (fundamental frequency, F0) in contrastive musical tones and timbre (first formant, F1) in contrastive speech vowels (Hutka, Bidelman, & Moreno, 2015). Both groups showed similar performance for F0, but musicians were better at discriminating F1. Musicians also showed enhanced MMN responses to both music and speech, implicating that speaking a tone language might not be associated with neural enhancements in this domain. These results corroborate the idea that the active engagement of cortical circuitry might depend on the cognitive relevance of the stimulus (Chandrasekaran, Krishnan, & Gandour, 2009; Bidelman, Gandour, & Krishnan, 2011ab). Arguably, musicians' expertise is gained through practice with a broader spectrum of pitch manipulations and productions within complex melodies, compared to the pitch patterns experienced by tonal language speakers during speech perception, which could explain these differences.

Another EEG study (Chandrasekaran, Krishnan, & Gandour, 2009) showed larger MMN responses in native Mandarin speakers than musicians. However, in this study, the researchers used iterated rippled noise with time varying F0 contours as stimuli, and additionally, modelled two of three stimuli after Mandarin Chinese tones T1 (high-level) and T2 (rising). Although the stimuli had no formant structure or temporal envelope, they comprised the energy bands for F0 and its harmonics that might have made them perceptually more similar to pitch contours that are familiar to Mandarin speakers. In fact, when comparing the third stimulus, that was not modelled after Mandarin tones, to T2, Chinese listeners were much less accurate, as they considered these two sounds to be one category. They also showed larger MMN responses to this difficult to distinguish contrast. Likewise, a study using pure tones varying in pitch height or interval distance demonstrated that Mandarin speakers were behaviorally more accurate and showed more neural sensitivity to small pitch changes and interval distances (Giuliano et al., 2011). Their ERP responses were not only earlier to relative change trials than to no change trials, but they also showed an earlier differentiation of trials by change direction when compared to non-tonal speakers that did not present with such sensitivity. However, as discussed earlier, non-native musicians represent these features equally well, suggesting that mere linguistic experience might not be the only explanation for the observed differences.

The presented evidence emphasizes the similarities between tonal language speakers and musicians, pointing to their shared expertise in pitch processing that might be responsible for their altered processing of pitch changes. However, Mandarin Chinese musicians demonstrate increased MMN response amplitude to lexical tone changes as well as faster behavioral discrimination performance compared to demographics-matched non-musicians, suggesting that enhancements driven by musicianship or speaking a tone language are not equivalent (Tang et al., 2016). These findings indicate that building on long-term L1 language experience, musicianship can additionally modulate the cortical plasticity of tone processing and is associated with enhanced neural processing of speech tones.

However, the influence of language experience on auditory processes is not limited to the auditory cortex. A large body of cross-language studies using frequency-following responses (FFR) has explored the efficiency of sound encoding within subcortical and cortical structures and showed pitch encoding enhancements arising from tonal language experience. Mandarin speakers have stronger pitch representations and exhibit smoother pitch tracking patterns than English speakers (Krishnan et al., 2005; Krishnan, Swaminathan, & Gandour, 2009). Greater and less variable neural phase-locking in Mandarin speakers is driven by sharpened response properties of neurons tuned to the lexical relevance of pitch contour in their L1. For example, Chinese listeners exhibit more robust pitch strength than English non-musicians, but only to stimuli with rapid pitch changes (Bidelman, Gandour, & Krishnan, 2011). This finding corroborates with previous FFR studies (e.g., Swaminathan, Krishnan, & Gandour, 2008; Krishnan, Swaminathan, & Gandour, 2009) claiming that the advantage of tone language experience does not uniformly apply to all aspects of pitch processing, but instead manifests itself within portions of F0 contour that exhibit linguistically relevant variations, likely reflecting experience with short-time changes at the syllable level. This is supported by experience-dependent enhancements of pitch encoding seen in the brainstem that seem to extend only to time-varying features of dynamic pitch patterns that native speakers of a language are exposed to (Swaminathan, Krishnan, & Gandour, 2008; Krishnan, Swaminathan, & Gandour, 2009). Taken together, these results suggest that neural patterns encoding acoustic information are not static, but instead might be modulated by L1 background (Krishnan et al., 2005) and possibly also plasticity due to L2 learning (e.g., Song, Skoe, Wong & Kraus, 2008).

1.3 Putative mechanisms behind perceptual strategies

So far, the discussed literature has provided evidence showing that a wide array of differences in speech perception or production strategies (e.g., Jasmin et al., 2021) and neural sound encoding (e.g., Krishnan et al., 2005) influence the way we listen to speech and learn new languages later in life. However, despite all these differences, speech perception strategies adjust to the unreliability of acoustic inputs (e.g., cochlear implant users; Winn & Chatterjee, 2012), difficulties in perceiving important acoustic cues (e.g., congenital amusia; Peretz, 2016), challenging listening conditions (e.g., masking noise, Gordon, Eberhardt, & Rueckl, 1993; overlapping speech, Symons, Holt, & Tierney, 2023), or foreign accents (Idemaru & Holt, 2011), and we learn new languages that feature entirely new sounds that do not exist in our first language. People do learn new languages even at an older age, surpassing the sensitive periods (Flege, 1987) or other constraints (e.g., auditory processing deficits, van Staden & Purcell, 2016; hearing problems; McConkey-Robbins, Green, & Waltzman, 2004) and some L2 learners achieve near-native performance (Birdsong & Molis, 2001). That means that the seemingly rigid perceptual strategies are adjustable (Jasmin et al., 2023) and modifiable by training (Lim & Holt, 2011).

Cross-linguistic influences

Due to continued L1 specialization in speech perception, acquiring non-native phonemes in adulthood might be a difficult task. Nevertheless, the ability to acquire new sound categories is maintained throughout the lifespan to some degree, although it is said to depend on the preexisting structure of L1. The degree of success in L2 learning seems to be related to several factors, namely the similarity of L2 to L1 (e.g., Bradlow, Pisoni, Akahane-

Yamada, & Tohkura, 1997), the degree of competition vs integration between native and non-native strategies (transfer vs interference, e.g., Best, 1994; Flege, 1995) and effective re-weighting of available acoustic information (e.g., Francis & Nusbaum, 2002). Indeed, listeners appear to dynamically adjust and selectively weigh various sources of information and acoustic cues during speech perception, and they vary considerably in their strategies and efficiency of doing so. Furthermore, the rate of learning and its final outcome is modulated by experiential variables such as the age of acquisition (AOA, e.g., Flege, Yeni-Komshian, & Liu, 1999) or length of residence in L2 environment (LOR, e.g., Granena & Long, 2012), and various cognitive abilities (musical skills, Chobert & Besson, 2013; auditory processing, Kachlicka, Saito, & Tierney, 2019; Sun, Saito, & Tierney, 2021; attentional switching, Kim & Hazan, 2010; inhibitory control, Kim, Clayards, & Kong, 2020; phonological encoding, Cutler, 2015a; auditory-motor integration, Shao, Saito, & Tierney, 2022; working memory, Miyake & Friedman, 1998; Darcy, Park, & Yang, 2015).

L1-L2 distance

Results from multiple studies demonstrate gradients of performance for non-native contrasts suggesting that some of the new contrasts might be easier to distinguish than others (e.g., perception of four Hindi voicing contrasts by English learners; Polka, 1991; identification of lenis-aspirated Korean voicing contrast by English learners, Kim & Hazan, 2010). These differences in difficulty appear to depend on the degree to which the native and non-native phonemes conflict with or are similar to one another, hinting at the possibility that relative distance between L1 and L2 phonological systems might have an effect on L2 speech attainment. Indeed, the influence of phonological distance on the ease and speed of language learning has been a topic of interest in second language acquisition research (e.g., Derakhshan & Karimi, 2015; Pajak, Fine, Kleinschmidt, & Jaeger, 2016; Georgiou, 2021ab). However, so far understanding of L1-L2 similarity effects is hindered by the lack of a coherent definition (e.g., phonological similarity, Bradlow et al., 1997; prosodic similarity, de Looze & Rauzy, 2011; Tremblay, Broersma, Coughlin, & Choi, 2016; Tremblay, Kim, Shin, & Cho, 2020; acoustic similarity, Mielke, 2012; Wu et al., 2021; lexical similarity, Holman et al., 2009) and an objective measure of linguistic distance (e.g., lexicostatistics, Bakker et al., 2009; Levenshtein distance, Petroni & Serva, 2010; neighborhood density, Stamer & Vitevitch, 2011).

In spite of these methodological inconsistencies, the similarity between the L1-L2 pair was shown to facilitate a positive transfer, i.e., the acquisition of new L2 contrasts without compromising their pronunciation. A large-scale study examined data from over 50000 people who completed a state-administered exam for L2 Dutch language proficiency (Schepens, van Hout, & Jaeger, 2020). A cumulative measure of phonological, morphological and lexical similarity was derived from comprehension and production data from L1 speakers of over 60 languages (e.g., German, French, Russian, Hindi, Vietnamese) speaking Dutch as a foreign language. Compared to the baseline model assuming the same level of difficulty across new categories, accounting for categorical similarities between the languages improved performance and explained more variance. Furthermore, similarity at the phonological, morphological and lexical levels appeared to be independent predictors explaining L2 learnability and combined together provided the best explanation of L2 success across domains. The predictive power of language distance was also demonstrated for the acquisition of subsequent languages in multilingual speakers (bilingual learners of Dutch; Schepens, van der Silk, & van Hout, 2016). Language background similarity was also an equally good predictor of L2 achievement as assessed via widely used measures of language aptitude (e.g., for Germanic vs non-Germanic languages; Bokander, 2020). At the same time, listeners can also fail to distinguish between L2 sounds or differentiate them from their L1 speech categories, providing evidence of negative transfer (or interference). Linguistic dissimilarity was also indicated as a factor contributing to age-related decline in adult language acquisition (Schepens, van Hout, & van der Silk, 2022).

L1 to L2 transfer and interference

These interactions between L1 and L2 language inventories can be partially explained by the Perceptual Assimilation Model (PAM, Best, 1993, 1994; its extension PAM-L2, Best & Tyler, 2007), which incorporates both contrastive phonological and noncontrastive phonetic influences from L1 addressed earlier by the Speech Learning Model (SLM; Flege, 1986, 1995) and Native Language Magnet Model (NLM; Kuhl, 1991; Kuhl et al., 1992; Iverson & Kuhl, 1995). All three models are built on the assumption that learning subsequent languages is strongly linked to the existing structure of L1. They posit that in contact with an unknown L2, listeners will consider new sounds as instances of L1 categories produced with varying degrees of similarity (i.e., better or worse examples thereof), or non-speech sounds if they do not resemble the units from their L1 system. However, these models take a slightly different stance when referring to underlying auditory mechanisms – pointing either to general (speech and non-speech processing utilizing the same auditory resources, e.g., NLM) or specialized mechanisms (linguistic-phonetic module uniquely for speech; e.g., Liberman et al., 1967, Liberman & Mattingly, 1989), and perceptual mechanisms operating on acoustic or articulatory information. It is also unclear to what degree these models converge on their predictions.

SLM is concerned primarily with the acquisition of phonological segments by experienced learners. It posits that new L2 sounds undergo equivalence classification against the existing L1 classes (Flege 1995), a process of comparing speech samples based on their phonetic similarity. The more experience learners have with the L2 (i.e., longer LOR; Flege & Liu, 2001) and the better the quality of that experience (i.e., more exposure to L2 and interactions with native speakers, MacKay, Flege, Piske, & Schirru, 2001; Flege & Liu, 2001), the more successful the formation of new sound categories. The SLM model predicts that the more dissimilar the L1 and L2 samples are, the better the expected L2 category configuration would be, suggesting that divergence between L1 and L2 categories can serve as an advantage in language learning. This model does not make any assumptions as to the speciality of auditory resources or level of information, but it posits that the capacity for the successful acquisition of new speech sounds is never completely lost.

On the other hand, NLM suggests an early-life acoustic prototype formation for L1 categories, the development of which relies on general mechanisms processing acoustic information (Kuhl, 2000). According to this approach, prototypical or best exemplars are clustered around the centre, creating the phonological space warped around its core with low within-class and excellent between-class discriminability around the prototype and the opposite pattern for the space near the categorical boundaries. This results in the predicted asymmetry in discrimination difficulty of prototypical vs non-prototypical stimuli (i.e., perceptual magnet effect). Due to a lack of appropriate experience with L2 acoustic inputs, listeners fail to develop clearly defined prototypes for L2 categories, resulting in more fuzzy representations (i.e., imprecise categories).

According to the PAM model, newly encountered non-native phones would be assimilated into the existing L1 system according to their perceived similarity to L1 categories. L2 tokens sounding like approximations of L1 classes would be categorized as L1 classes, although how well they fit within these classes might vary. L2 phones that could fit into two or more native categories would remain uncategorized, whereas others that are completely different from L1 sounds would not be assimilated at all and might be considered as non-speech. Further classification of assimilation types divides them into two-category assimilation (each L2 sound assimilated to a separate L1 category), single-category assimilation (both L2 sounds assimilated to the same L1 category, equally good representations of L1 class), category goodness difference (both L2 sounds assimilated to the same L1 category, but one is a better exemplar of L1 than the other), uncategorized-categorized (one L2 sound assimilated to L1 category, second L2 sound within unfamiliar phonetic space), or uncategorized-uncategorized (both L2 sounds within unfamiliar phonetic space), and non-assimilable (non-speech; Best, 1993). We can consider this taxonomy of cross-language contrasts as a framework for understanding why some contrasts are easy to learn while others are almost impossible to acquire. Such a model provides a range of successful predictions. For example, research demonstrated focalized (similar to one L1 category), clustered (similar to multiple L1 categories) and dispersed (not similar to any L1 category) uncategorized assimilation types (Egyptian Arabic speakers' perception of Australian English; Faris, Best, & Tyler, 2016) and excellent discrimination along two-category assimilation, medium to good discrimination of categories differing in the goodness of fit, and the worst performance on single category assimilation (English speaker's perception of Zulu and Tigrinya; Best, McRoberts, & Goodell, 2001), both consistent with the PAM's prediction of phonetic space gradience. However, some research demonstrated that it is unclear whether all the phonemes follow these predictions (Tyler, Best, Faber, & Levitt, 2014). It is also important to note that these predictions apply only to the initial contact with new L2 categories. Based on the principles of perceptual learning, these predictions could be extended to account for changes that are taking place during prolonged L2 exposure (Best & Tyler, 2007) and learning in a classroom-settings (Tyler, 2019).

The perceptual interference account (Iverson et al., 2003) extends the concepts of perceptual nonuniformity proposed by Kuhl (1991) and offers a more detailed explanation of how perceptual warping of acoustic space affects speech perception. It states that the early language experience irrevocably alters low-level perceptual processing and that these changes interfere with the formation and adaptability of higher-level linguistic representations in adulthood. One of the most widely researched examples in language learning is Japanese speakers' acquisition of the English /r/ vs /l/ contrast. Japanese native speakers lack sensitivity to third formant (F3) differences along the English /r/ vs /l/ boundary, making it extremely difficult to learn. Japanese speakers eventually improve their perception of this contrast, but these improvements might be achieved by using other more easily perceivable cues (Ingvalson, McClelland, & Holt, 2011). Indeed, research showed that Japanese listeners rely on F2 when resolving the /r/ vs /l/ English contrast or integrate information from F2 and F3 (Yamada & Tohkura, 1992). The effect of this mistuning of L1 and L2 perceptual spaces on language learning is twofold. First, since people rely on salient cues (i.e., more easily perceivable) that are not necessarily the most reliable, this likely leads to more categorization errors. Secondly, listeners' increased sensitivity to irrelevant acoustic dimensions might place a strain on processing resources that might be crucial factors in determining L2 learning success. For example, detecting critical acoustic differences might require more focused attention and longer processing times.

Cue weighting adjustments in attention-based models

As a way to address some shortcomings of perceptual assimilation and interference models alike, an idea of dynamic cue re-weighting during speech categorization was put forward (e.g., Francis, Baldwin, & Nusbaum, 2000; Francis & Nusbaum, 2002). The early attention-to-dimension models (A2D) were based on the General Context Model, perceptual learning and categorization (Nosofsky, 1986; Pisoni, Lively, & Logan, 1994; Goldstone, 1994) derived directly from the classical Prototype (Rosch, 1973, 1975) and Exemplar theories

(Nosofsky, 1991, 1992) concerned with issues of organizing conceptual knowledge with classification, similarity and probability density estimates. The A2D models assume a spatial representation of perceptual space defined by a range of acoustic dimensions. The dimensional warping of that space[3] is operationalized here in terms of attentional operations and formalized as weights or multipliers assigned to acoustic dimension (Holt et al., 2018). Depending on attentional focus these weights stretch or shrink the perceptual space (Francis & Nusbaum, 2002). For example, focusing attention on a given dimension increases its weight and stretches it so that the differences along that dimension are more easily perceivable. According to these models, the set of available dimensions describing the speech stimuli is somewhat constant for each listener. What is changing as a function of L1 experience, L2 learning or training, is the relative weight assigned to existing dimensions. Increased or decreased weight emphasizes an important piece of information or downplays the role of irrelevant or otherwise non-informative dimensions (Francis & Nusbaum, 2002). Such mechanism can explain the re-structurization of phonological space observed in the course of language learning (e.g., Bradlow et al., 1997; Bradlow & Pisoni, 1999; McCandliss et al., 2002; Iverson, Hazan, & Bannister, 2005; Kondaurova & Francis, 2010).

As noted earlier, listeners enter their L2 speech learning journey with a pre-existing specialization for categorizing native sounds. That implies that the initial representation of perceptual space is defined in terms of dimensions especially useful for a given language. Learning a new language, therefore, might involve shifting attention away from dimensions that are no longer useful and towards the dimensions that can help resolve the new speech contrasts (Francis, Baldwin, & Nusbaum, 2000). This reorganization of the phonological space in response to attentional shifts is reflected in spatial stretching and shrinking of these dimensions. However, acquiring the phonology of another language might involve

[3] This idea is somewhat similar to what Kuhl and collaborators (Kuhl, 1991; Kuhl et al., 1992; Iverson & Kuhl, 1996) proposed in their Native Language Magnet Model, but the warping described there was localized in nature (i.e., describing shrinkage nearby phonemic category centres).

more elaborate operations than a simple redirection of attention from one cue to another. For some contrasts, it might be possible to achieve a good category separation by increasing attention to the underrated dimension while decreasing attention to the unnecessary cue (Francis, Baldwin & Nusbaum, 2000). But the acquisition of other categories might necessitate cue integration or separation of cues that in L1 are mutually reinforcing. It is also possible that listeners might not redistribute their attention if a combination of other available dimensions would allow them to successfully perform the task (i.e., changes take place only in cases when it benefits the listener; Francis & Nusbaum, 2002). Furthermore, if L2 learners lack the ability to detect the crucial cue, they might also use cues that are secondary for native speakers (e.g., using duration to distinguish between English vowels /i:/ vs /i/ by speakers of various L1s; Polish; Bogacka, 2004; Russian; Kondaurova & Francis, 2008; Mandarin; Flege, Bohn, & Jang, 1997; using F0 in English voicing decisions by Spanish speakers; Llanos, Dmitrieva, Shultz, & Francis, 2013). Some flexibility in cue use is also necessary to accommodate for the consistent perception of linguistic contrasts in different contexts (e.g., Chambers et al., 2017) and across various speakers and accents (talker normalization; Nusbaum & Morin, 1992; Nusbaum & Magnuson, 1997). However, recent research has shown that normalization and perceptual learning might operate on distinct levels of processing (Lehet & Holt, 2020).

Supporting the application of A2D models to perceptual learning, Francis and Nusbaum (2002) showed that improvements in category learning arise from acquiring within-category similarity and between-category distinctiveness. If the two L2 tokens are sufficiently distinct from one another, then focusing on the within-class similarities might be a sufficient strategy to learn to distinguish them. However, when L2 categories are difficult to tell apart, using the same set of cues might not suffice. In that case, to expand the perceivable distance between the L2 categories, listeners might need to learn to use a new cue or pull one from an integral acoustic blend by readjusting attention assigned to those unused dimensions. This process would eventually stretch their perceptual space to accommodate new L2 categories. Such interpretation could be somewhat problematic because it assumes an innate and limited set of cues accessible to each listener, rejecting the possibility of learning new acoustic dimensions. However, if we consider acoustic cues as features processed by our auditory system and represented in the brain (e.g., Cohen et al., 1999), apart from the physiological limitations of our auditory system to process these signals there is no evidence demonstrating limitations to learning a completely new linguistic contrast (e.g., whistling, Meyer, Dentel, & Meunier, 2017; clicking, May & Werker, 2014). As we already discussed, even though the auditory system tunes in to L1 sounds very early on and makes perceiving cues in other languages more difficult, we do not lose the ability to hear various acoustic properties in a non-linguistic context or improve auditory sensitivity to these properties with targeted training.

All these processes suggest that speech perception learning does not depend on passive transformations of acoustic signals into perceptual representations or pattern-matching process that links stable linguistic representations with auditory properties (for discussion about passive vs active processes see Heald & Nusbaum, 2014). Passive processing would imply that none of the input is modulated in any way or poses any constraints on cognitive resources, which as we already discussed is not true during speech perception. More recent views on speech perception learning try to incorporate influences of context and experience, and they do that by introducing active processes of attentional modulation, neural plasticity, and feedback into their theoretical frameworks (e.g., Davis & Johnsrude, 2007). It is also possible that perceptual weights during speech perception are impacted by the basic auditory representations at the early stages of processing since, as shown earlier, some dimensions are more robustly encoded than others (e.g., stronger pitch signature for tonal language speakers; Swaminathan, Krishnan, & Gandour, 2008; Krishnan, Swaminathan, & Gandour, 2009; Bidelman, Gandour, & Krishnan, 2011).

Learning mechanisms in speech perception

Speech perception learning, therefore, is not a simple feature-category mapping problem. Apart from combining visual, lexical, syntactic, and contextual information, L2 speech acquisition involves detecting, reweighting, and integrating information from various acoustic features and at the same forming new categories by dynamically mapping these cues, separating categories from existing ones, or forming entirely new representations. Such a complex learning process is said to be supported by dual learning systems (DLS;

e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2005). Inspired by research in the visual domain[4], dual learning comprises two systems – an explicit reflective system and an implicit reflexive system. Reflective learning is an explicit hypothesis-testing system that involves top-down processes (Ashby et al., 1998). This type of learning seems to be optimal for the acquisition of so-called rule-based categories that can be easily described with verbalized rules. Such rules are easy to define when categories can be separated based on one or two dimensions (see Roark & Holt, 2019 where for simplicity of the experimental setup, they used a rule-based category with only one dimension), but rule formulation can get complicated very quickly when considering multidimensional perceptual spaces with interdependent dimensions, like in the case of speech perception. On the other hand, the reflexive system involves implicit procedural learning that is optimal for learning "information-integration" categories that require integration across at least two dimensions. These approaches presume that the development and consolidation of rules for classification would lead to automatized or more implicit processing.

Some researchers proposed that speech category learning is dominated by reflexive learning (e.g., Chandrasekaran, Yi, Smayda, & Maddox, 2016). Their main argument was that linguistic categories are defined based on rules that are not only very difficult to verbalize, but also can be defined by multiple highly variable and redundant acoustic dimensions (e.g., VOT contrast can be defined by 16 acoustic features; Lisker, 1986) that prohibit a simple dimension-to-category mapping. In one study, Roark and Holt (2019)

[4] Although interpreting visual theories in the auditory domain seems plausible, it is not straightforward. The dimensions on which we operate in the auditory domain are much more difficult to capture and describe as the number of words dedicated exclusively to describing auditory sensations is somewhat limited and accessible mostly to sound experts (e.g., musicians, acousticians, sound engineers; Alias, Socoro, & Sevillano, 2016). For example, terms like frequency modulation, formant or brightness usually require some explanation to naïve listeners. On the other hand, perceptual dimensions widely used in the visual domain, such as shape, direction, or colour are easily understood and perceived by most people. This might mean that some of the assumptions of the dual systems model might not directly translate to the auditory domain. These differences might also suggest that balancing the difficulty between dimensions while selecting stimuli for experiments might not be an easy task.

claimed that easily separable dimensions might serve rule-based learning, whereas those that are not as clear would likely lead to information integration. They showed that the integration strategies persisted even if they were suboptimal for completing the task. People also had a problem disengaging from irrelevant dimensions and were biased towards integrating dimensions that are positively correlated, suggesting that they might indeed be too difficult to tell apart. Further research corroborated this idea by showing enhanced category learning during training including procedures facilitating reflexive learning (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014). Explicit instruction has also been shown to be beneficial in reflexive-optimal tasks (Chandrasekaran et al., 2016). However, not all the linguistic contrasts are acquired due to cue integration. Recent work also showed that there are no evident differences in acquiring rule-based vs information-integration categories and that the task performance relies mostly on learners' strategies (Roark & Chandrasekaran, 2023).

Basic principles of perceptual learning

Perceptual learning operates by implicitly associating perception with actions that lead to representation reinforcement. However, this strengthening is not based on a simple association of learned stimuli with the desired response but rather on learning to attend to the cues that will facilitate required discrimination (Sutherland & Macintosh, 1971). In a broad sense, perceptual learning can be understood as learning that leads to changes in perceived stimulus characteristics that happen through repeated exposure to that stimulus (Gibson, 1963). It includes gradual improvements in speed (e.g., faster detection of shifted elements; Ahissar & Hochstein, 1997) or accuracy of perception (e.g., Bradlow et al., 1997) or development of new perceptual representations or processes (Lehet, Fenn, & Nusbaum, 2020). In the context of speech perception, learning involves generating new L2 speech categories that can then be quantified by successful identification and discrimination of phonemic boundaries. We assume that category formation reflects either creation of a new representation or some quantitative changes in representations existing prior to learning via attentional shifts or decreased or increased salience. However, the processes underlying learning are rarely studied, and in most cases, the main focus is given to learning outcomes.

The literature described several important aspects of perceptual learning (Shiffrin & Lightfoot, 1997). Firstly, perceptual learning is said to change the subjective perception of some stimuli. For example, language learners were shown to shift their strategies and interpretations depending on which language they thought they were listening to (Yazawa, Whang, Kondo, & Escudero, 2020). Secondly, the observed changes do not occur spontaneously but rather reflect listeners' experience with stimuli characteristics. These changes can be almost immediate (e.g., sine-wave speech learning, Cheng, Xu, Gold, & Smith, 2021) or more gradual and arise as a function of the extended exposure to stimuli characteristics. Cue weighting in L2 perception seems to be closely related to learners' L2 experience and proficiency level, supporting the idea that perceptual strategies gradually change throughout learning. Indeed, more experienced learners are better at inhibiting and enhancing relevant acoustic cues (Escudero, Benders, & Lipski, 2009; Kong & Yoon, 2013; Kong & Edwards, 2015), but longitudinal evidence of these changes is yet to be seen. Third, learning can be described in terms of stages or depth of processing. Early stages would involve basic sensory analysis and then learning would progressively advance in complexity. Presumably, the observed changes would also depend on a given phase of learning or depend on a training regime. However, a recent study showed similar gains and generalization effects across a range of experiments differing in stimulus variability, sequencing and instruction types (Obasih et al., 2023). Finally, successful learning requires some form of automatization of the processes involved in perception to allow their efficient use in everyday communication.

Ahissar (1999) extended the definition of perceptual learning by providing insights into how we can map behavioral improvements to the understanding of underlying cortical areas. The specificity of observed training improvements suggests the involvement of early sensory cortical areas, whereas the role of complex attentional modulations suggests involvement of high-level cortical areas. This proposal is in line with the precedence of learning – easy cases are learned first, and difficult cases later. One possible explanation of such a pattern is that learning easy examples is more general than learning difficult ones, which might require the involvement of more specialized brain networks. The reverse-hierarchy model (Ahissar & Hochstein, 1997) argues that learning proceeds in reverse order to sensory processing. That means it begins from higher-level representations that are usually easily accessible to learners (e.g., sentences or words they wish to understand or distinguish), towards more fine-grained lower-level processing that is concerned with specific stimulus details (e.g., encoding of pitch contour or formant frequencies). Typically, higher-level representations can be accessed and attended to at the later stages of processing as they directly represent objects in our real-life environment.

Assuming such hierarchical nature of processing, learning of easier elements needs to happen first to facilitate learning of more complex examples. This is because difficult examples build on the preexisting knowledge of easier elements and is solved by recruiting higher-level areas. As a consequence, training under challenging conditions leads to more specific learning (i.e., limited to higher-level specific representations) than training under easy conditions. Indeed, we often observe that training outcomes do not transfer well to novel conditions, even if they differ from the original training conditions only in simple attributes (e.g., Loebach & Pisoni, 2008). Only in some cases, participants successfully transfer gains from learning material to new situations, which might be related to the training framework (e.g., cue enhancement vs cue inhibition; Kondaurova & Francis, 2010; but see Obasih et al., 2023 for no evidence of difference across various training regimes). Overall, transfer along a continuum of difficulty seems to be an effective training procedure (Sutherland & Macintosh, 1971). Training failures might occur due to the incompatibility of difficulty levels with the direction of learning pathways that hinder learning transfer. Since difficult cases are meant to be learned with more resolution, it might not be possible to generalize this knowledge to novel stimuli characteristics as these might be represented by different groups of specialized low-level neurons (but see Ahissar, 1999 where a single informative presentation was sufficient to enable learning under difficult conditions).

We can conclude that perceptual learning of speech might involve shifting attention between acoustic cues to reflect changes in the importance of these features in L2. Most research to date relies in their inferences on observations of behavioral differences in cue weighting between people coming from different L1 backgrounds, having various degrees of expertise, or measuring performance before and after a brief perceptual training, but no concrete evidence of mechanisms behind those changes was offered. This raises the question of whether such differences in cue weighting are reflected in attentional shifts towards more important or salient L2 cues or their increased salience and whether these strategies can be adjusted as a result of perceptual training. There is little evidence showing how L2 learners distribute their attention during their L2 learning or characterizing the mechanisms of such transitions (Nusbaum & Goodman, 1994). Also, so far, research investigating the effects of perceptual training has been mostly focused on simple artificial stimuli (e.g., Liu & Holt, 2011) or phonemic contrasts (e.g., /beat/ vs /bit/; Ylinen et al., 2010; /beer/ vs /pier/; Idemaru & Holt, 2011). Under such conditions, category formation is greatly simplified and depends on the limited selection of dimensions of interest picked for comparison. Although not ecologically valid, such stimuli allowed researchers to understand the behavioral changes governing perceptual learning. However, any generalizations about real-life language learning conditions should be treated with caution. More research tracking the development of perceptual strategies in L2 learning in more naturalistic conditions is needed to fully illuminate this issue.

1.4 Attention as a scaffold for L2 learning

Two classes of theories have been discussed in an attempt to explain possible interactions between L1 perceptual weighting strategies and their realizations in L2. Proponents of the first view claim that L1 language speech categories interfere with L2 speech categories (Iverson & Kuhl, 1994; Flege, 1995; Best et al., 2001), so some effects of that interference between L1 and L2 phonological inventories would be seen. Another explanation, the dimension-selective attention account, proposes the existence of a "perceptual strategies profile". Depending on the individual's L1 background, we can expect them to have their cue preference biased in one way or another since they learned to value the salience of specific signals important in their L1 more than any other signals from subsequent L2 (Francis, Baldwin, & Nusbaum, 2000; Francis & Nusbaum, 2002). This also means that available acoustic cues such as pitch, duration and amplitude are weighted according to that profile so that more informative or statistically reliable cues will receive more attention (Holt et al., 2018). The hypothesis of the crucial role of selective attention was supported by results showing that people might alter their strategies depending on the temporary changes in their usefulness (Idemaru & Holt, 2011) or when they are distracted (Gordon et al., 1993; Symons, Holt, & Tierney, 2023). Furthermore, the dimension-selective attention model predicts that greater experience in pitch would change general pitch sensitivity (Jasmin et al., 2020) or that impairments in this domain will also be transferred (Jiang et al., 2010). These findings suggest that learning an L1 can change the salience of different acoustic dimensions, shaping the perceptual strategies that develop as people learn new categories. Learning an L2 might require people to direct attention to perceptual dimensions which they have grown used to neglecting.

Attention in SLA research

Despite the fact that redirecting attention and changes in salience have been repeatedly mentioned as an underlying mechanism of L2 speech learning, little evidence was presented to support these claims. Second language acquisition (SLA) research made limited attempts to investigate the role of attention in L2 learning. However, so far, attention has been considered mainly at the level of general executive functions or cognitive abilities needed for allocating resources necessary to process language inputs (attentional control, e.g., Segalowitz & Frenkiel-Fishman, 2005; Darcy, Park & Yang, 2015; inhibition, e.g., Blumenfield & Marian, 2013). Attention understood as a cognitive ability is of particular interest in SLA research. Better attention and more flexible inhibition have a positive impact on the effectiveness of L2 speech perception and the overall facilitatory effect on language learning. Previous research demonstrated links between attention and speech perception, but attention was measured at a more global level (e.g., Test of Everyday Attention; Ou & Law, 2017), which is not directly linked to auditory attention or attention to acoustic dimensions that are relevant in the context of speech perception and learning, but rather more general input manipulation abilities.

A more focused operationalisation of attention appears in the study by Segalowitz & Frenkiel-Fishman (2005). The researchers talk about "attention control" defined as the cost of shift during the linguistic task-switching paradigm (participants classified time adverbials and causal connectives displayed in predictably changing quadrants of the screen). They argued that more proficient learners should be more adept at managing their attentional resources while performing such a task and found that attention control explained 59% of variance in L2 proficiency of English French bilinguals they tested. When they controlled for L1 effects, L2-unique attention control explained 32% of variance. In another study, Darcy, Park and Yang (2015) talked more specifically about "selective attention" as a process of selecting relevant dimensions during linguistic decision-making and "attentional flexibility or control" describing dimensional switching, even though the operationalisation of dimensions focused on comparing different sources of information, namely voice identity (male vs female voice) and lexical information (word vs non-word) rather than acoustic properties of speech. Although they hypothesised that attention control would be related to individual differences in phonological processing in a way that more efficient attentional control would facilitate bringing important information to focus, their results did not support these claims. Inhibitory control is also believed to be linked to L2 proficiency. An eye-tracking study measured crosslinguistic coactivation during word recognition and showed that more proficient L2 speakers presented more parallel language activation and less Stroop effects than less proficient English learners of Spanish (Blumenfield & Marian, 2013). Another eye-tracking study showed that increased inhibitory control is linked to less competition between languages for English French bilinguals in spoken word processing (Mercier, Pivneva, & Titone, 2013). These results suggest that crosslinguistic competition influences domain-general inhibition.

It has also been argued that language learning is hindered because learners use their L1 strategies that are no longer adequate for L2. For example, Japanese listeners tend to rely more on F2 information than F3 when distinguishing between English /r/ and /l/ (Iverson et al., 2003) even though F3 is a primary cue marking that contrast and F2 is only useful for recognizing /l/. Similarly, Chinese listeners rely more on pitch when categorizing English lexical stress than on other cues like duration, intensity, or vowel quality (Yu, & Andruski, 2010; Zhang & Francis, 2010). These discrepancies in cue weighting might be a driver of inaccurate L2 speech perception. The ability to selectively attend to relevant acoustic dimensions has been suggested to drive the adjustment of cue weighting patterns (e.g., Francis et al., 2000; Francis & Nusbaum, 2000) and linked to increased L2 proficiency. For example, results pointed to the potential role of attentional shifts in language learning by showing changes in relative weights of F0 vs VOT depending on the listeners' L2 proficiency (English learners of Korean, Kong & Edwards, 2015; Korean learners of English, Kong & Kang, 2022), but no direct evidence of attentional enhancement was presented in these studies. Only recently, SLA research started considering dimension-selective attention as one of the possible predictors of L2 success (e.g., Saito et al., *under revision*), focusing on its potential role in successful perception of various L2 contrasts.

It seems that the lack of coherent theory and explanation of the proposed mechanisms arises partially due to the inconsistent understanding of attention in the literature. Various approaches conceptualized attention in terms of perceptual warping or attentional shifts towards more informative or reliable cues to distinguish subtle differences in sentence prosody, word pronunciation or individual phonemes. Yet, the majority of research to date has focused on a more general understanding of attention, with only the most recent efforts attempting to address the issue of redirecting attention within the sound (i.e., attention to acoustic dimension; Holt et al., 2018; attention to particular values along a specific acoustic dimension, Laffere, Dick, & Tierney, 2020; Laffere, Dick, Holt, & Tierney, 2021) rather than between streams (i.e., attention to an auditory object, e.g., Bergman, 1990; Alain, 2007; Shinn-Cunningham, 2008; for review see Lee, Larson, Maddox, & Shinn-Cunningham, 2014).

Attention-modulated cortical tracking of acoustic dimensions

Selective attention directs neural resources to the relevant aspects of speech signals to extract the information needed to understand it (e.g., selecting specific speakers from overlapping talkers, focusing on the relevant acoustic cues that define linguistic categories). While the exact processes involved in attentional modulation are not fully explained, studies have consistently shown that selective attention can modulate the cortical representations of auditory events by enhancing those that are in the focus of attention and suppressing the representations of the competing sources (e.g., Chait, de Cheveigné, Poeppel, & Simon, 2011). For example, attention-driven enhancements of cortical tracking have been demonstrated for attended compared to ignored speech (e.g., Viswanathan, Bharadwaj, & Shinn-Cunningham, 2019; Reetzke, Gnanateja, &

Chandrasekaran, 2021) or tone sequences (Elhilali, Xiang, Shamma, & Simon, 2005; Laffere et al., 2020, 2021). In the context of speech perception learning, we are interested in attention guided towards specific dimensions within the auditory objects (i.e., dimension-selective attention; Holt et al., 2018) or particular values along a specific acoustic dimension (i.e., region-specific selective attention, Nosofsky & Hu, 2022), as theories of speech perception suggested that re-allocating attention depending on cue informativeness or reliability plays a role in cue weighting (e.g., Gordon et al., 1993; Francis et al., 2000; Francis & Nusbaum, 2002; Holt et al., 2018).

One way to measure dimension-selective attention to specific acoustic dimensions within a single stream is to look at its neural signatures while participants are asked to attentively track changes along a specific acoustic characteristic of the presented stimuli. In an fMRI study, participants listened to two streams of low (250 Hz) and high (4000 Hz) streams of tones presented simultaneously but to different ears (Da Costa et al., 2013). The sequences were series of two-interval forced choice trials presented one after another, each containing 5-6 tones. Participants were asked to identify whether the subsequent sequence was the same as the preceding one. They were also asked to switch attention from low to high frequency every 30 s. Results showed that spectrally selective attention leads to enhanced neural response to frequency-tuned voxels within the attended frequency range, as evidenced by the increased blood-oxygen-level-dependent (BOLD) activation. These results are consistent with an earlier study in which participants detected irregularities in the rhythm of the attended stream (Paltoglou, Sumner, & Hall, 2009). They were presented with low (223-281 Hz) and high frequency (3564-4490 Hz) streams and instructed to attend to one of them in subsequent blocks. Similarly to Da Costa et al.'s study, results demonstrated frequency-specific attentional enhancement to the attended frequency range, although with less detailed frequency mapping (1.5 x 1.5 mm x 2.5 mm compared to 1.5 mm isotropic resolution). Dick et al. (2017) further corroborated these results by showing that attentionally driven maps in each hemisphere are similar to the detailed tonotopic maps within the primary auditory areas. Some efforts to track attention to different features such as pitch and sound location have also been made, showing stronger activation in the premotor and supplementary motor cortex related to attention to location and no area specific to attention to pitch (consistent with the suggested auditory “what” and “where” processing streams, Degerman et al., 2006). However, activations specific to attention to pitch were also found. For example, compared to talker selection, attention to talker pitch resulted in bilateral but right-dominant activation in the superior temporal sulcus (Hill & Miller, 2010) and attention to target pitch range in greater activation in the left posterior superior temporal sulcus associated with pitch categorization (Lee et al., 2013).

One disadvantage of brain imaging is that typically MRI signal intensity (i.e., BOLD activation) is rather sluggish as it takes around 6 s from the stimulus onset to develop and another 20 s to return to baseline (Jenkinson & Chappell, 2018), which is much slower than within syllable acoustic changes that must be tracked in speech perception (between 3.3 and 5.9 syllables per second in English; Arnfield & Roach, 1995). Understanding the dynamics of auditory attention might require different analyses, and MEG and EEG are particularly well suited to provide measurements with high temporal resolution. Early EEG research found that responses in the primary auditory cortex as early as within 100 ms from the sound onset are sensitive to attentional modulation (i.e., larger N1 for attended compared to unattended stream; Hillyard et al., 1973). Another study found that also mismatch negativity (MMN) response was larger for attended stream in the dichotic listening task of tone sequences (Woldorff & Hillyard, 1991), even though it was previously thought of as an independent of attention stimulus-driven feature detecting system (Naatanen et al., 1988).

Exogenous vs endogenous attention

Auditory attention can be controlled by top-down factors, such as cognitive selection of information, knowledge, expectations and current goals, and bottom-up factors that reflect behaviourally relevant sensory stimulation or salience of a stimulus and its properties. Although it has been suggested that salience and selective attention can both enhance neural representations of auditory objects (Lee et al., 2014), these processes could be supported by two separate networks or mechanisms. Indeed, prior work has demonstrated that both the mismatch negativity (MMN) and P3 responses, associated with detection of acoustic change and orientation of attention respectively, are sensitive to the magnitude of the change along multiple acoustic dimensions (e.g., pitch deviants; Berti, Roeber, & Schroger, 2004; Escera, Alho, Winkler, & Naatanen, 1998; increasing amplitude of distractors; Rinne et al., 2006; changes in intensity, duration, pitch or timbre of trumpet, saxophone or clarinet notes presented in musical sequences, Shuai & Elhiali, 2014). A similar salience effect was observed during listening to stressed syllables – oddballs elicited increased P3a response, suggesting that in unattended speech, prosodic features capture listeners' attention (interpreted as a correlate of prosodic salience; Wang, Friedman, Ritter, & Bersick, 2005). Further research suggests that the effects of salience can be modulated by language experience. In an MMN study, English and Mandarin speakers listened to English nonwords with contrastive stress. Results showed that stress deviants elicited MMNs in both groups. However, English speakers' responses were stronger than those of Mandarin speakers, suggesting more robust and precise processing of English stress intensity patterns by native compared to non-native speakers (Chung & Bidelman, 2016).

If neural activity can track variations along different acoustic dimensions, we would observe attention-enhanced representations of attended dimensions and potentially suppressed representations of unattended dimensions. Indeed, studies using the frequency tagging paradigm showed that tracking changes along multiple stimulus characteristics presents as an enhanced response to the attended property. This paradigm was developed to quantify attentional modulation of neural responses to visual stimuli by measuring potentials elicited through the presentation of stimuli with distinctive flicker frequencies (Toffanin et al., 2009). But since then, tagging stimuli at different presentation rates has been adapted for tagging changes within sounds (i.e., acoustic dimensions can be targets of attention) and successfully applied to tracking competing speech streams (Bharadwaj et al., 2014), linguistic structures (Ding et al., 2016) or neural entrainment to beat and meter (Nozaradan, Peretz, Missal, & Mouraux, 2011).

More recently, research showed that frequency tagging could be used to measure not only dimension-selective attention but also dimensional salience as both modulate cortical tracking of various acoustic dimensions (duration and intensity, Costa-Faidella, Sussman, & Escera, 2017; pitch and spectral peak, Symons, Dick, & Tierney, 2021). Symons and colleagues (2021) manipulated the salience of acoustic dimensions by altering pitch differences between the presented tones hypothesising that larger pitch differences would be perceived as more salient and therefore elicit a stronger response. Results demonstrated that cortical tracking was indeed modulated by the increased pitch salience – stronger cortical tracking of pitch changes was shown for sequences with a 2-semitone pitch difference compared to the condition with a 1-semitone difference.

1.5 Open questions

Although a substantial amount of research has been undertaken to understand L2 acquisition, we still know relatively little about how various factors shape listening strategies, what consequences these strategies might have for learning subsequent languages in adulthood, and what might be the mechanisms driving these changes (or lack thereof) during learning. The purpose of the present research is to answer several lingering questions about perceptual strategies underlying second language acquisition.

First, I asked whether native language experience and musical training would lead to differences in cue weighting strategies geared toward the most informative acoustic dimensions. Building on the reviewed theories of speech perception (Francis & Nusbaum, 2002), I hypothesized that lifetime experience with L1, as well as years of targeted training would shift listeners' attention toward these dimensions and enhance their salience. I predicted that listeners would up-weight the dimension that is particularly relevant in their L1 (i.e., pitch for conveying meaning in Mandarin Chinese) or was acquired during training (i.e., pitch for conveying musical structures for trained musicians) and explored whether these strategies extend to other domains (i.e., speech vs musical beats). I also tested whether the differences in cue weighting strategies are linked to enhanced dimension-selective attention or dimensional salience of relevant acoustic dimensions.

Second, to understand the sources of individual differences in cue weighting, I investigated whether and, if so, to what extent language learning experience, dimension-selective attention and dimensional salience can explain the observed differences in cue weighting strategies. In accordance with the cue weighting theories, I expected that attention, salience, or both would emerge as significant predictors of cue weighting strategies. Furthermore, I examined whether these strategies are stable across tasks.

Third, I looked into the roles of dimension-selective attention and dimensional salience in L2 learning. To that end, I investigated to what extent language learning experience, dimension-selective attention and dimensional salience can explain the variability in L2 learning outcomes. To provide a more in-depth understanding of how these factors might contribute to language proficiency, I examined these predictors in the context of L2 segmental and suprasegmental perception, suprasegmental production, vocabulary and grammar knowledge.

Finally, I asked whether adjusting perceptual strategies with perceptual training is possible. I designed a targeted training to redirect listeners' attention away from their preferred dimension (i.e., pitch) towards another acoustic dimension (i.e., duration). I measured their behavioural cue weighting patterns, prosody perception, ability to attend to pitch and duration and their dimensional salience before and after the training. I attempted to answer the question of what are the consequences of such training on participants' cue weighting strategies, their ability to selectively attend to acoustic dimensions, the relative salience of these dimensions, and whether changes along any of those measures translate to better L2 prosody perception. Given that the perceptual strategies employed in people's L1 might interfere with the acquisition of new L2 strategies, I measured participants' performance and their strategies immediately after the training and after 6 months to see whether any changes in these strategies would be maintained over time.

To answer these questions, I conducted a series of cross-sectional and longitudinal experiments using behavioural and neuroimaging methods. An overview of the research design is presented in Figure 1.

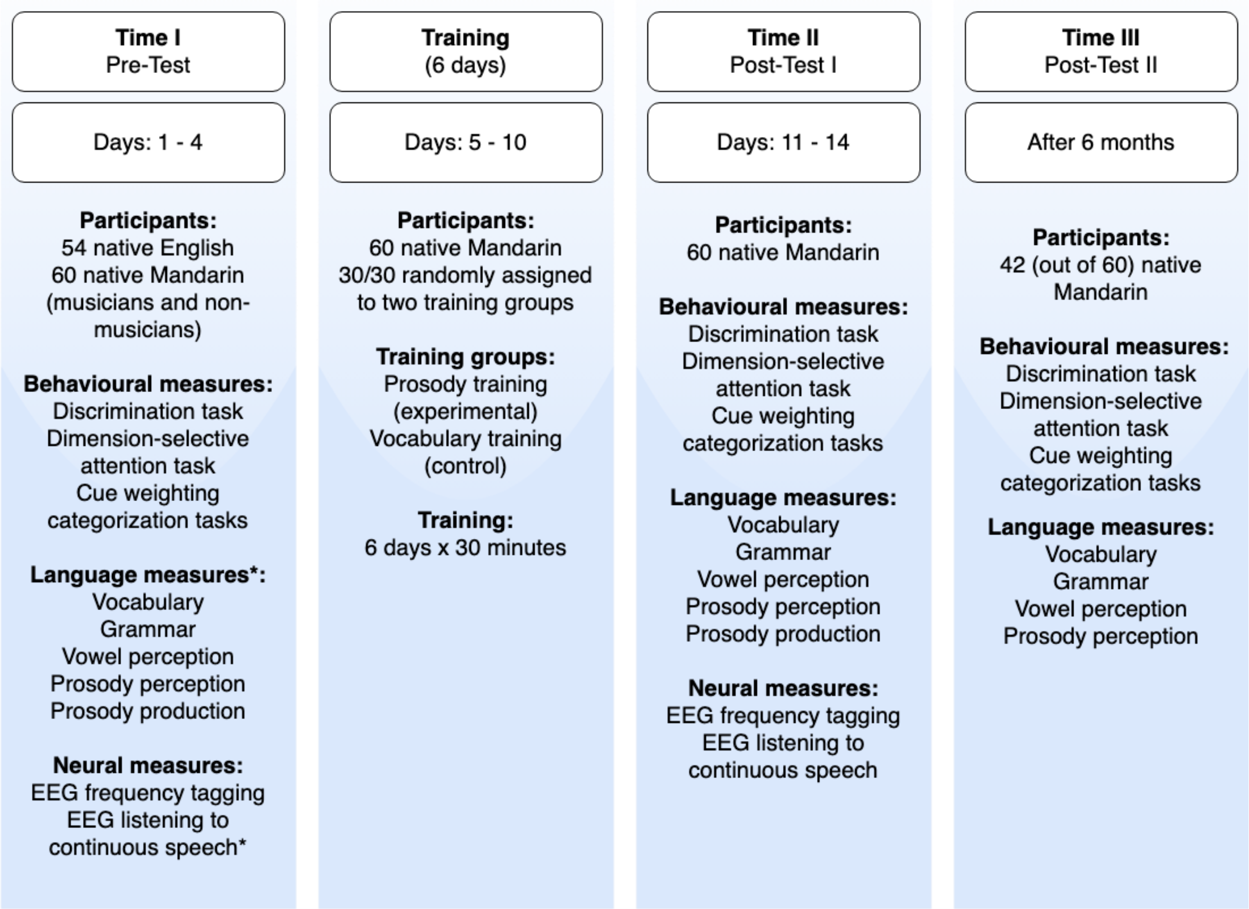


Figure 1. Overview of the research design. I recruited two groups of participants, native English and Mandarin speakers. English speakers completed assessments only at Time I and I collected their data on behavioural tasks and neural measures (excluding tasks marked with an asterisk – language measures and EEG listening to continuous speech were collected only for Mandarin speakers). In Chapter 2, I compare the behavioural performance and neural data of native English and Mandarin speakers at Time I. Chapters 3 and 4 delve in a more detailed discussion of factors supporting L2 learning, based on Mandarin speakers' data at Time I. In Chapter 5, I present the training paradigm, and compare the dimension-selective attention, dimensional salience and cue weighting patterns of Mandarin speakers before and after the training (i.e., Time I and Time II).

Thesis outline

Chapter 2 is dedicated to reporting empirical work comparing the perceptual strategies employed by speakers of two different languages, English and Mandarin Chinese (English and Mandarin speakers' behavioural and neural data at Time I). It provides a detailed report of results from EEG and behavioral tasks that compared auditory processing, dimension-selective attention, dimensional salience, and cue weighting strategies between tonal and non-tonal language speakers with and without musical training. This chapter answers whether L1 background and musical training influence speech perception and whether these strategies extend to non-speech stimuli. The following two chapters are concerned with perceptual strategies employed by Mandarin Chinese learners of English and report results from EEG and behavioral studies (Mandarin speakers' data at Time I). Chapter 3 investigates individual differences amongst Mandarin speakers with a specific focus on their speech perception and production and cue weighting strategies, whereas Chapter 4 investigates predictors of L2 learning success among Mandarin Chinese learners of English. I used regression models to predict second language success defined by different aspects of L2 language performance: speech prosody perception and production, but also vocabulary and grammar knowledge. The discussion concentrates on the importance of various factors in language learning and attempts to reconcile the knowledge from second language acquisition research with new insights from neuroscience. The second-to-last chapter is devoted to the prosody training paradigm designed for the purpose of this project. Chapter 5 reports the short-term training effects across a range of neural and behavioural measures and discusses the potential applications and drawbacks of the presented paradigm (Mandarin speakers' behavioural and neural data at Time I and Time II). It also discusses the long-term effects after 6 months of immersion in an L2 English environment (Mandarin speakers' behavioural data at Times I, II, and III). The thesis closes with a general discussion and conclusions presented in Chapter 6. The thesis outline is depicted in Figure 2.

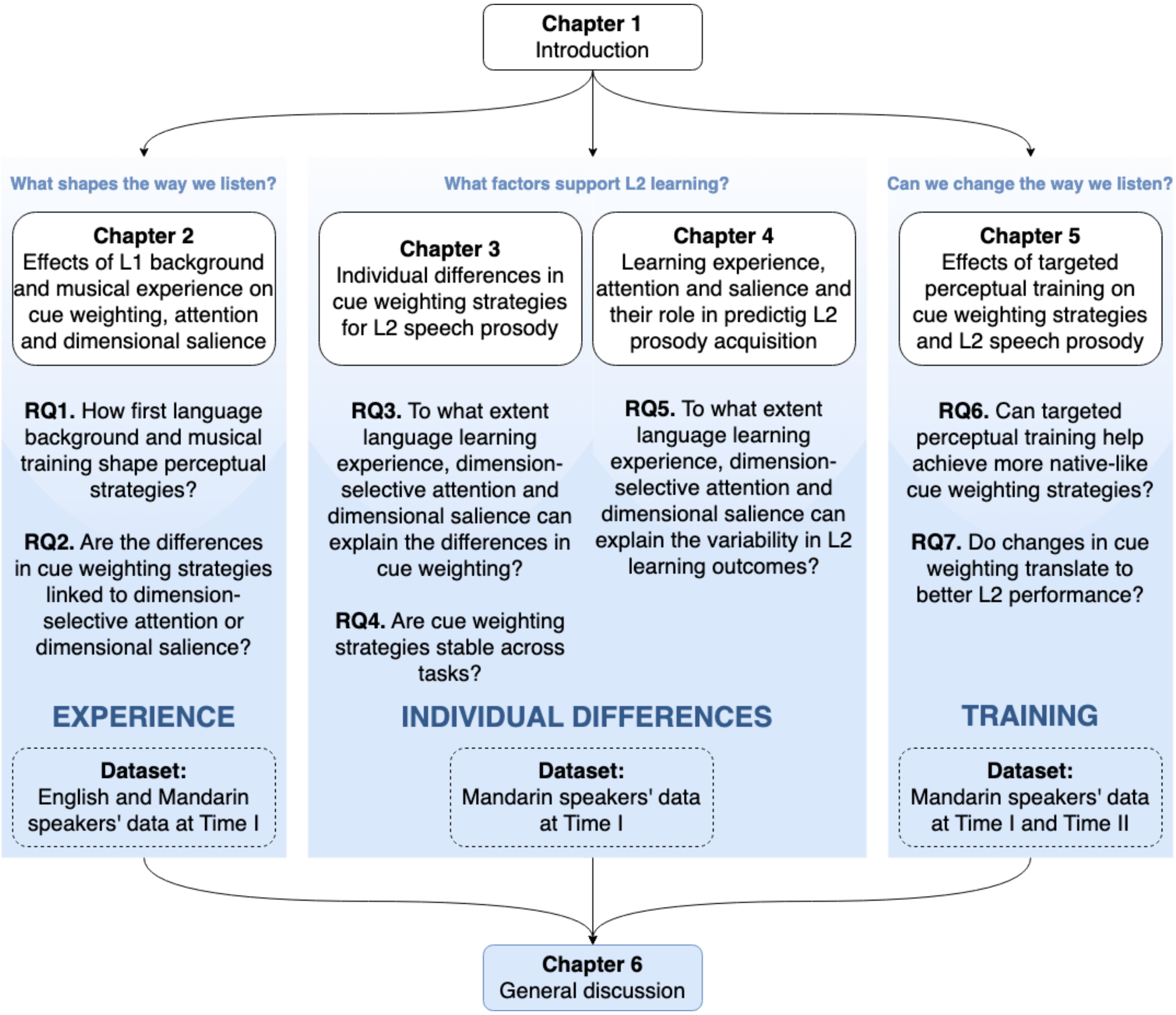

Chapter 1
Introduction
What shapes the way we listen?
What factors support L2 learning?
Can we change the way we listen?
Chapter 2
Effects of L1 background and musical experience on cue weighting, attention and dimensional salience
Chapter 3
Individual differences in cue weighting strategies for L2 speech prosody
Chapter 4
Learning experience, attention and salience and their role in predictig L2 prosody acquisition
Chapter 5
Effects of targeted perceptual training on cue weighting strategies and L2 speech prosody
RQ1. How first language background and musical training shape perceptual strategies?
RQ2. Are the differences in cue weighting strategies linked to dimension-selective attention or dimensional salience?
RQ3. To what extent language learning experience, dimension-selective attention and dimensional salience can explain the differences in cue weighting?
RQ4. Are cue weighting strategies stable across tasks?
RQ5. To what extent language learning experience, dimension-selective attention and dimensional salience can explain the variability in L2 learning outcomes?
RQ6. Can targeted perceptual training help achieve more native-like cue weighting strategies?
RQ7. Do changes in cue weighting translate to better L2 performance?
EXPERIENCE
INDIVIDUAL DIFFERENCES
TRAINING
Dataset:
English and Mandarin speakers' data at Time I
Dataset:
Mandarin speakers' data at Time I
Dataset:
Mandarin speakers' data at Time I and Time II
Chapter 6
General discussion

Figure 2. Thesis outline.

Chapter 2. Effects of first language background and musical experience on cue weighting, attention, and dimensional salience

Abstract. While categorizing speech, listeners use perceptual strategies that allow them to select relevant information and place sufficient importance on its source. Research suggests that not all available information receives equal attention and that these strategies are driven by the increased salience of acoustic dimensions reliable in speakers' native language (L1). According to attention-based theories, learning a new language might involve shifting attention away from L1-relevant dimensions and towards dimensions that are more informative in a new language. Musical training also involves attending to specific acoustic dimensions, but whether it would have consequences for language learning is unknown. To date, the role of attention and dimensional salience in these processes has not been sufficiently tested. In this study, I set to answer whether there are any systematic differences in how L1 background and musical training influence cue-weighting strategies in music and speech and if L1 and musical experience are linked to one's ability to selectively attend to acoustic dimensions or enhanced salience of these dimensions. A group of 54 native English and 60 Mandarin Chinese learners of English completed a set of behavioural tasks to assess their auditory thresholds in perceiving differences in pitch and duration, dimension-selective attention to these dimensions, and cue weighting strategies in categorizing English prosody and musical beats. Using a frequency tagging EEG paradigm, I also measured dimensional salience of pitch and duration. Mandarin speakers, compared to English speakers, showed enhanced attention to and preferential use of pitch across behavioural tasks (up-weighting pitch during prosody and musical beats categorisation and demonstrating superior attention to pitch in verbal and non-verbal stimuli). However, there was no effect of language background on neural entrainment to acoustic dimensions. Although tone language speakers benefit from an enhanced ability to direct endogenous attention to pitch when it is task-relevant, they do not experience increased involuntary exogenous capture of attention by pitch. Comparison of cue weighting strategies between participants with and without musical training revealed that musicianship sharpens tuning to a task-relevant dimension. These results are consistent with attention-to-dimension theories of cue weighting which claim that listeners redirect their attention toward the most informative or task-relevant cues.

2.1 Introduction

Prior research suggests that first language (L1) background plays a significant role in shaping perceptual strategies underlying second language (L2) acquisition. The similarity between phonological inventories (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997) or prosodic patterns between native and foreign language (de Looze & Rauzy, 2011; Tremblay, Broersma, Coughlin, & Choi, 2016; Tremblay, Kim, Shin, & Cho, 2020) have been found to contribute to the difficulty of learning a new language such that the lower the similarity between L1 and a given L2, the more challenging it would be to learn that language. These difficulties arise partly due to the fact that lifetime experience with sounds in our L1 makes us experts in distinguishing various phonemes from one another (e.g., Kuhl et al., 1992) and weighting acoustic dimensions according to their reliability in predicting category membership in that language (Francis, Baldwin, & Nusbaum, 2000;

Toscano & McMurray, 2010). This specialization in processing L1-relevant information comes at the price of difficulties experienced while learning a second language later in life – listeners struggle with detecting new contrasts that differ from or do not exist in their own language. Because various linguistic contrasts can be described with multiple acoustic cues that will not be equally important or easily perceivable by speakers of different languages, listeners might weigh acoustic information embedded in speech in a number of ways.

A classic example of these difficulties is that of the contrast between /r/ and /l/. This contrast distinguishes many minimal pairs in English, but because the Japanese language has a single consonant that, in terms of third formant (F3), corresponds to English /l/ (Hattori & Iverson, 2009), the contrast between /r/ vs /l/ is very difficult to acquire for native Japanese speakers (e.g., Yamada & Tohkura, 1991; Bradlow et al., 1997). It seems that Japanese speakers do not show enough sensitivity to detect differences in F3 crucial for resolving the contrast between the two and they tend to rely on F2 instead (Iverson et al., 2003; Ingvalson, McClelland, & Holt, 2011). In the same way, native English speakers have not learned to distinguish between various ways of using relative pitch height differences or changing pitch contour, which, on the other hand, are crucial for conveying lexical and sub-lexical information in tonal languages (e.g., Francis, Ciocca, Ma, & Fenn, 2008; Hao, 2018). Non-tonal languages like English make use of pitch in a more secondary role signalling linguistic focus (greater pitch movement for emphasized phrase; e.g., Breen et al., 2010), emotional state (various patterns of pitch levels and contour types linked to different emotions; e.g., Rodero, 2011), speakers' intentions (e.g., exaggerated pitch patterns for sarcasm and irony; Attardo, Eisterhold, Hay, & Poggi, 2003; increased pitch during attempted deception; Streeter et al., 1977) or sentence type (e.g., rising pitch contour for questions; Bartels, 1999), while other cues redundantly convey lexical and sub-lexical information. These discrepancies in the relative importance of various acoustic cues across languages might mean that an optimal L1 listening strategy will not be as effective for any other subsequent L2.

A lifetime of experience with L1-specific dimensions tunes the auditory system to acoustic characteristics of that language that carry relevant information (Holt & Lotto, 2006). Research provides abundant evidence that such extensive experience can change how we process sound, emphasizing, for example, tonal language speakers' expertise in processing pitch. To perceive lexical tones, listeners must integrate pitch information across speakers and articulatory and intonational information across utterances, and these two aspects of pitch (namely, speaker-normalized pitch height and pitch changes) are said to be processed independently. Recent findings from electrocorticographic recordings (Li et al., 2021) revealed that speaker-normalized pitch is not language-specific, as it can be used, to a lesser extent, by English speakers when processing sentence prosody and intonation. On the other hand, Mandarin speakers showed different temporal response function (TRF) tuning curves for pitch features, reflecting the differences between the two languages (i.e., wider pitch range and lower relative pitch). As evidenced by studies using frequency-following responses (FFR), they have stronger pitch representations and exhibit smoother pitch tracking patterns than English speakers (Krishnan, Xu, Gandour, & Cariani, 2005; Swaminathan, Krishnan, & Gandour, 2008; Krishnan, Swaminathan, & Gandour, 2009; Bidelman, Gandour, & Krishnan, 2011). Mandarin native speakers were also more precise in pitch discrimination (e.g., Giuliano et al., 2011), pitch interval discrimination (quantified as the relative pitch difference between two consecutive pitches; e.g., Pfordresher & Brown, 2009; Hove, Sutherland, & Krumhansl, 2010; Giuliano et al., 2011; Creel et al., 2018; Zheng & Samuel, 2018) and had better melody discrimination (Bradley, 2012; Bidelman et al., 2013; Wong et al., 2012; Liu, Hilton, Bergelson, & Mehr, 2023).

It is hypothesized that extensive expertise in perceiving pitch variations and attending more to pitch changes than to changes in other less-informative dimensions enhances its relative salience. Indeed, Mandarin Chinese speakers were shown to rely more on pitch information when perceiving English speech and musical beats, were better at attending to pitch and ignoring amplitude, and had difficulties ignoring pitch even when explicitly asked to do so (Jasmin et al., 2021). They place more importance on pitch and less on other acoustic information while listening to English stress (Wang, 2008; Yu & Andruski, 2010; Zhang & Francis, 2010) and phrase boundaries (Jasmin et al., 2021; Zhang, 2012), and even overuse pitch in speech production (Nguyen et al., 2008; Zhang, Nissen, & Francis, 2008). Existing attention-to-dimension theories of speech perception suggest that dimensions that are important for conveying structure in an individual's L1 become particularly salient or likely to capture attention (Francis & Nusbaum, 2002; Gordon, Eberhardt, & Rueckl, 1993; Holt et al., 2018*)*. According to these models, this acquired salience of the language-specific dimensions might lead to upweighting of that dimension during perception and drive increased attentional gain to that dimension. However, this hypothesis has not been systematically tested and very little empirical evidence of these processes was ever provided to support these claims.

The OPERA hypothesis (Patel, 2014ab) suggests musical experience as another influential agent instigating changes to the auditory system since it also involves learning to selectively attend to specific acoustic dimensions. Learning music might place higher demands on the auditory system than speech perception because speech understanding requires much lower pitch encoding precision and is more robust to pitch perturbations (Patel et al., 2010) than distinguishing pitch movements within a musical sequence, where even a one semitone difference might be structurally relevant (Eerola et al., 2006). Indeed, musicians were shown to outperform non-musicians on a range of tasks related to pitch perception (e.g., pitch discrimination, Micheyl, Delhommeau, Perrot, & Oxenham, 2006; detection of frequency modulation, Carey et al., 2015; perception of Cantonese tones by English speakers; Choi, 2020). Further supporting the idea of musical training heightening perceptual acuity to acoustic characteristics, Symons & Tierney (2023) demonstrated that musical experience is linked to enhanced attention to relevant acoustic dimensions and ability to ignore irrelevant dimensions in speech stimuli and reliance on a single primary dimension during suprasegmental categorization (duration for phrase boundary and pitch for linguistic focus contrasts). These findings hint at the possibility that musical experience allows for refining listening strategies that, as a consequence, are more flexible and use the most informative dimension that would help them to perform the task effectively. On the other hand, an ERP study (Hansen et al., 2022) presented evidence of more integrative neural processing of acoustic information across musically relevant stimuli by subjects with musical experience. Authors posited that the specialized refinements in predictive processing might enable musicians to capitalize upon complex domain-relevant acoustic cues in a flexible fashion. However, it must be noted that both studies focused on acoustic dimensions embedded in entirely different stimuli (i.e., speech vs musical tones) therefore their conclusions are not easily comparable. Taken together, these findings indicate that up-weighting dimensions that learners mastered throughout their life history in L1 or musical training and down-weighing dimensions that are unfamiliar or more difficult to perceive leads to differential strategies from native speakers or untrained listeners.

Distinctive patterns of responses to various aspects of pitch are also seen when comparing the performance of tonal language speakers to musicians who do not speak a tonal language. Tone language speakers and musicians both have shown enhanced neural encoding of pitch-relevant information at both subcortical (pitch contour or musical chords; Bidelman, Gandour, & Krishnan, 2011ab) and cortical (pitch contour; Chandrasekaran, Krishnan, & Gandour, 2009) levels, but the differential benefits for tonal language speakers and musicians without tonal language experience emerged when considering more complex aspects of pitch and music perception. While musicians' expertise is gained through practice with a broader spectrum of pitch variations within complex melodies, tonal language speakers may not have been exposed to such a wide variety of pitch patterns (e.g., four tone contours in Mandarin Chinese; van de Weijer & Sloos, 2014), which might contribute to the observed differences. Current research emphasizes the similarities between tonal language speakers and musicians, pointing to their shared pitch expertise that might be responsible for their altered pitch processing patterns. However, Mandarin Chinese musicians demonstrated increased MMN response amplitude to lexical tone changes as well as faster behavioral discrimination performance compared to demographically-matched Mandarin non-musicians, suggesting that pitch processing enhancements related to L1 background and musical training are not equivalent or perhaps are independent (Tang et al., 2016). These findings indicate that building on the long-term L1 language experience, musicianship can still modulate the cortical plasticity of tone processing and is associated with enhanced neural processing of speech tones. Taken together, these results suggest that neural responses to sound might be modulated by L1 background (Krishnan, Xu, Gandour, & Cariani, 2005), second language-learning related plasticity (e.g., Song, Skoe, Wong & Kraus, 2008), and musical training (Tang et al., 2016).

Present study

This study investigates the effects of language background and musical experience on perceptual categorization strategies, dimension-selective attention and dimensional salience. I hypothesized that depending on the L1 background and musical expertise, participants would weigh cues differently while performing speech categorization tasks. For example, they might not rely on the primary acoustic cue defining a particular linguistic category but on the dimension with which they are more familiar, and they might revert to secondary acoustic cues if the primary cue is not salient enough or reliable. More specifically, I predicted that Mandarin speakers would rely more on pitch while categorizing contrasts primarily defined by duration, such as phrase boundaries (as previously shown by Jasmin, Sun, & Tierney, 2021) and show stronger pitch reliance while categorizing tokens varying in linguistic focus and lexical stress. I test these hypotheses in the context of pitch and duration – pitch is a highly relevant cue in Mandarin Chinese but has secondary importance in English; duration was chosen as a dimension orthogonal to pitch in speech. Similarly to prosody contrasts, the location of musical beats within musical sequences can also be conveyed by placing emphasis on the musical notes. The note on which the beat falls is extended and shows a changed pitch contour (Ellis & Jones, 2009; Hannon, Snuder, Eerola, & Krumhansl, 2004). Given this partial overlap of relevant cues signaling structural differences in speech and music, I hypothesized that the effects of language experience might not be limited to particular language tasks and extend to other domains, possibly changing more general listening strategies. Including verbal and non-verbal stimuli and broader prosodic contexts allows me not only to replicate the results from previous studies, but also to address the question of whether the influence of language experience extends across domains, possibly changing more general listening strategies.

I also asked whether L1 and musical background are linked to one's ability to selectively attend to various acoustic dimensions. I hypothesized that better discrimination and attention-to-dimension abilities might be helpful while using various acoustic dimensions and adopting successful listening strategies. Therefore, these abilities might be tuned to cues that are especially relevant or reliable in participants' L1. In line with previous research, I expected Mandarin speakers to have better pitch discrimination thresholds and perform better than English speakers at attending to pitch. I also expected Mandarin participants to experience difficulty ignoring pitch while attending to duration in the dimension-selective attention task (Jasmin et al., 2021) and musicians to outperform non-musicians on pitch discrimination and attention tasks (Micheyl et al., 2006). Investigating both L1 background and musical training allows to test the interaction between these two types of experiences and test whether they have different effects depending on participants' L1.

Finally, I offer a unique insight into the putative mechanisms underlying speech perception strategies by testing whether years of L1 experience and musical training enhance the salience of relevant dimensions. To depart from the existing methods requiring participants' behavioural ratings (e.g., Kaya & Elhiali, 2014), I used the EEG frequency tagging paradigm to more objectively measure dimensional salience. Tagging stimuli at different presentation rates has been successfully applied to tracking competing speech streams (Bharadwaj et al., 2014), linguistic structures (Ding et al., 2016) or neural entrainment to beat and meter (Nozaradan et al., 2011). Recent research showed that frequency tagging can also be used to measure dimensional salience and dimension-selective attention – both modulate cortical tracking of various acoustic dimensions (duration and intensity, Costa-Faidella et al., 2017; pitch and spectral peak, Symons, Dick, & Tierney, 2021). My experimental setup allows me to test the hypotheses about the role of salience in theoretical accounts of auditory categorization. Following the previous literature (e.g., Liu et al., 2020), I predict that Mandarin speakers will show stronger pitch tracking (stronger pitch salience) in verbal and non-verbal stimuli.

2.2 Methods

Participants

Participants were recruited from advertisement platforms, social media outlets and via word of mouth. The group of English native speakers comprised students from various university departments (e.g., psychology, language sciences, education) recruited mainly from the SONA platform and professional musicians recruited from music job boards and social media groups for artists based in London. Mandarin speakers were students from various departments (e.g., psychology, education, economics, law, architecture, engineering) recruited from the SONA platform and social media community groups and societies (Facebook and WeChat). Since I did not know the effect size of interest, I could not estimate the sample size a priori. The sample size was motivated by the previous studies showing that a similar sample size was sufficient to detect sizeable effects (e.g., effect sizes of $A_{Vargha\text{-}Delaney}$=.49-.82 for comparisons of normalized cue weights between $N_{English}$=50, $N_{Mandarin}$=50, and $N_{Spanish}$=30; Jasmin, Sun, & Tierney, 2021), and limited by the maximum number of participants that could be recruited and tested given time and resource constraints. The aim was to recruit 60 native English and Mandarin speakers (30 with and without musical training in each group). A total of 61 English and 75 Mandarin speakers completed the study, however only the data from 54 English and 60 Mandarin speakers were included in the analyses. Participants who in the categorization tasks showed either a significant negative correlation between either stimulus dimension and categorization responses, or no significant relationship between either stimulus dimension or categorization responses (patterns suggestive of misunderstanding task instructions) were flagged for removal. Five Mandarin speakers were excluded based on their poor performance in the dimension-selective attention tasks (i.e., participants who did not achieve a minimum of 75% correct responses in the single dimension training blocks after three attempts), and ten based on their responses in categorization tasks. Six English speakers were excluded based on their responses in categorization tasks and one due to technical issues in the lab that prevented the researcher from recording their EEG data.

The majority of English participants were monolingual native speakers. Only 5 of them indicated speaking another language since birth (one bilingual English-Farsi, one English-Hungarian, one English-Russian, and two English-Bengali speakers), whereas 28 studied at least one other language starting from teenage years to early adulthood (e.g., Spanish, German, French, Portuguese, Hebrew, Russian, Italian). None of the participants had previous experience with tonal languages. Following the criteria described by Zhang et al. (2020), I considered as musicians only the participants who reported more than 6 years of systematic musical training (N=29). Although the 6-year cut-off is somewhat arbitrary, it captures the differences between the groups – non-musicians in this study had minimal to no musical training compared to musicians who all reported more than 6 years of training and practice playing various instruments. Most English musicians reported playing more than one instrument (only 6 played one instrument and 3 were professional singers). Most of them played either guitar or piano (N=15 for each instrument) and the rest played a large variety of other instruments (bass, clarinet, drums, violin, flute, trumpet, harp, oboe, recorder, cello, horn, bassoon or accordion). 9 English participants reported practicing music in their childhood (N_{4years}=1, N_{3years}=2, N_{2years}=3, N_{1year}=3), but admitted that they are no longer able to play any instruments. The remaining participants had no musical training.

Mandarin speakers were all speaking English as a second language but were not raised bilingually – they learned English at school and reported only 1 to 17 months (M=7.41, SD=3.21) of residence in English-speaking countries. I intentionally recruited very beginner L2 learners (recruitment criterion LOR <18 months), to compare their naïve phonetic representations to those of native speakers. 7 participants reported speaking an additional language (1 x Russian, 1 x French, 1 x German, 2 x Japanese and 2 x Korean). 29 Mandarin-speaking participants reported more than 6 years of musical training (Zhang et al., 2020), compared to non-musicians who had no music experience. Most participants with musical training reported playing piano (N=15; the remaining participants played various instruments such as violin, pipe, flute, guitar, bass or clarinet and trained in singing) and five participants were trained to play traditional Chinese instruments. Ten participants reported playing more than one instrument. A summary of the demographic information about both groups is displayed in Table 1.

Table 1. Summary of participants' demographic, language, and musical training background information.

		English (N=54)	***Mandarin[5] (N=60)***
Age	*Range*	18–38	18–31
	Mean (SD)	23.94 (5.62)	22.62 (3.27)
Gender	*Female*	37	53
	Male	17	6
	Non-conforming	0	1
Musical training (> 6 years)	*Yes*	29	29
	No	25	31
L2 experience	*Speaks L2 English*	NA	N=60
	Speaks other L2	N=28	N=7

Behavioural measures

Dimension-selective attention task

This task was designed to measure participants' ability to pay attention to changes along one acoustic dimension while ignoring changes in another dimension.

Stimuli

The base stimuli were eight unique tokens, four verbal and four non-verbal, varying along fundamental frequency (F0, correlate of voice pitch) and duration (Table 2). To create this 2x2 grid of discrete values, I extracted vowels from speech excerpts and generated acoustically matched tones. The speech stimuli were extracted from the phrase "Tom likes barbecue chicken" taken from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick & Tierney, 2020). I used two versions of this phrase, with and without emphasis placed on the word "barbecue" and extracted the first vowel /a/ from both versions to capture clearly audible natural within-vowel pitch and duration variations occurring in English speech (the selection of the word "barbecue" was somewhat arbitrary). Both stimuli were ramped with a 10-ms on/off cosine ramps before processing to avoid acoustic transients. To create pitch-varying stimuli, I morphed the emphasized and

[5] Additional information about Mandarin speakers' history of L2 English learning and use was also collected, but it's beyond the scope of this chapter and will be discussed in subsequent chapters.

non-emphasized vowels along the F0 dimension using the STRAIGHT voice morphing software (Kawahara & Irino, 2005). The morphing procedure involved extraction of the F0 from voiced parts of the recordings and analysis of periodic aspects and filter characteristics of the signal. The final step included manual marking of corresponding salient portions of the recordings (i.e., anchor points), based on which I generated 100 morphed samples that acoustically represented a smooth transition of F0 values from the emphasized to non-emphasized vowels (i.e., level 100 contained F0 information suggesting emphasis placed on a vowel, whereas level 1 the vowel without emphasis). The duration and other acoustic parameters were kept constant. I selected two samples that differed from each other by approximately 2 semitones (levels 1 and 56; difference = 2.03 semitones) to make the differences easily perceivable. Then, I used Praat software (Boersma & Weenink, 2001) and a custom script (Winn, 2014) to morph the duration of the vowel to 70.58 and 175.83 ms (difference = 105 ms) and created a 2 (pitch) x 2 (duration) stimulus grid using the selected stimuli. There are small differences in F0 (<0.5 Hz) between the different duration levels, but these are not perceptible. The reason for selecting those values was to keep the differences in F0 and duration balanced for pitch vs duration salience. The base non-verbal stimulus units consisted of complex tones with 4 harmonics and were generated to match acoustic properties of verbal sounds (ramped to avoid perception of transients). The tones varied along two dimensions: pitch (F0) and duration. There were 2 levels of pitch and duration, resulting in 4 unique tokens for each domain (Table 2).

Table 2. Base stimuli parameters. Four unique sounds in each domain were created as the combination of pitch and duration levels (i.e., pitch1_duration1, pitch1_duration2, pitch2_duration1 and pitch2_duration2). These sounds were used as base units for creating sequences for the dimension-selective attention task, frequency tagging paradigm, and generating continua for the discrimination tasks.

Domain		***Pitch (mean F0, Hz)***	***Duration (ms)***
Verbal (vowel /a/)	Level 1	110.88	70.58
	Level 2	124.40	175.83
Non-verbal (tones)	Level 1	110.88	70
	Level 2	124.72	175

Stimuli sequences

The individual verbal and non-verbal tokens were then concatenated into 2 Hz sequences in which pitch and duration varied at different rates (every 3 sounds = 0.67 Hz and every 2 sounds = 1 Hz). Repetitions, or instances where the dimension did not vary at the expected rate, were inserted into half of the sequences for each dimension. This resulted in 4 trial types: pitch repetition only, duration repetition only, repetitions in both dimensions, and no repetitions in either dimension. Thus, the stimuli in each domain and attention condition were identical, varying only in the focus of attention. From each stimulus set (verbal and non-verbal), 64 stimuli (32 stimuli varying in pitch every 2 sounds and in duration every 3 sounds and 32 stimuli varying in duration every 3 sounds and in pitch every 2 sounds) were randomly selected and assigned to either attend pitch or attend duration conditions (32 trials per attention condition). The stimuli were assigned to the opposite attention conditions in two versions of the task to counterbalance items across subjects (versions A and B). For example, if a given recording with pitch changing every 2 sounds and duration every 3 sounds were assigned to attend pitch in version A, it would be assigned to attend duration in version B. Participants were randomly assigned to complete Version A or B of the task.

Task

Participants listened to sequences of verbal and non-verbal sounds changing in pitch and duration at two different rates. At the beginning of each block, they were asked to pay attention to changes in one of the acoustic dimensions. Each trial began with 500 ms of silence, followed by the presentation of the stimulus. Once the stimulus had finished playing, text appeared on the screen asking participants whether they heard a repetition within the attended dimension. Participants responded by clicking the 'Yes' or 'No' button on the screen. Feedback was provided on each trial. Participants received the next set of instructions between blocks and could take a break.

Prior to the task, participants listened to examples of different pitch and duration levels and sequences where only a single dimension was changing. Participants then completed a short training task with these sequences. The training task was blocked by attention conditions but with the rate of the attended dimension randomized. At the start of each block, participants were informed which dimension to attend to and the rate at which that dimension was expected to vary. Participants received 8 trials per attention condition (4 per rate) in which only the attended dimension varied. Participants were required to answer at least 6 out of 8 trials (75%) correctly on each training module to move on to the next task. If participants failed to reach the performance threshold, they could repeat the training for that dimension up to 3 times, and they were not allowed to continue to the next stage of the study if they failed to do so.

The trials in the main task were identical to the training task, but here both dimensions were changing. For each stimuli type (verbal and non-verbal), the task consisted of 4 blocks presented in a random order, each corresponding to a different condition (2 attention conditions x 2 rates of change). At the start of each block, participants were told which dimension to attend to and the rate at which that dimension was expected to vary. Participants' responses were recorded and the hit rate (collapsed across dimension change rate) for each dimension was computed as the dependent variables.

Auditory discrimination tasks

All participants completed a series of discrimination tasks to assess their discrimination thresholds for verbal and non-verbal stimuli along two acoustic dimensions, pitch and duration (i.e., 4 tasks in total). Tasks were presented in a random order, which was balanced across participants with a Latin square group assignment function on Gorilla.

Stimuli

For both verbal and non-verbal stimuli, I created 100-step continuums of stimuli along pitch and duration dimensions, with the other dimension held constant. Having selected the endpoints of the pitch continuum for verbal stimuli (levels 1 and 56; see the detailed description of stimulus creation in the "Dimension-selective attention task" section), I used STRAIGHT voice morphing software to create a 100-step pitch continuum between those two endpoints while holding duration constant at 105 ms. To create duration discrimination stimuli, I used Praat software to create a 100-step duration continuum by changing the unstressed version of the vowel from 70 ms to 175 ms while keeping their pitch constant. The non-verbal stimuli consisted of complex tones (4 harmonics with 10-ms linear on and off ramps). Complex tone pitch and duration continuums were created with F0 and duration matching that of the speech stimuli (see Table 2). The pitch discrimination continuum consisted of 100 stimuli with F0 varying from 110.88 Hz to 124.40 Hz while duration was held constant at 105 ms. The duration discrimination continuum consisted of 100 stimuli varying from 75 ms to 175 ms with a constant F0 of 110.88 Hz.

Procedure

In each trial, participants were presented with three stimuli in AXB format, with X matching either stimulus A or B. In each test, participants heard three sounds, for example – speech sounds varying in duration presented with a constant interstimulus interval of 500 ms – and needed to decide whether the first or third sound was different from the other two by pressing the appropriate number (1 or 3) on the screen. I used an adaptive three-alternative forced-choice procedure modified from the transformed up-down procedure described by Levitt (1971). The task difficulty increased after every second correct response and decreased after every incorrect response. The presentation began at stimulus level 50 with an initial step size of 10. That meant that the task became easier or more difficult by 10 steps. However, after a first reversal, the step size changed to five, after a second reversal to two, and after a third reversal to one and remained at this level until the end of the presentation, allowing me to detect the smallest difference between the stimuli participants can hear. Each task stopped either after 70 trials or seven reversals. The final score was calculated as the levels of each reversal from the second onward (i.e., the lower the threshold level, the better performance).

Prosodic cue weighting tasks

Stimuli

All the speech stimuli were taken from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick & Tierney, 2020) and included sentences meant to capture contrasts in three prosody features: linguistic focus, phrase boundary, and lexical stress (Table 3) and musical beat patterns. The speech tokens were created by recording the voice of a native Southern British English speaker reading all phrases listed in Table 3. Identical portions of the recordings (i.e., "study music", "If Barbara gives up", and "compound") were then extracted, and two versions (Token A and Token B) morphed together using STRAIGHT software (Kawahara & Irino, 2005; Jasmin et al., 2020) by adjusting the values of F0 and durational morphing rates orthogonally to create the stimuli.

Table 3. Prosodic cue-weighting tasks' target phrases and words. Stimuli were derived from recordings of two contrastive sentences or words (capitalization indicates contrastive focus). Linguistic focus stimuli derived from sentences "Dave likes to STUDY music, but he doesn't like to PLAY music" and "Dave likes to study MUSIC, but he doesn't like to study HISTORY", phrase boundary stimuli from sentences "If Barbara gives up, the ship will be plundered" and "If Barbara gives up the ship, it'll be plundered", and lexical stress stimuli from a word "compound" pronounced with stress placed on the first or second syllable. The identical portions of each recording were extracted to obtain two versions of the same phrase that differed in the location of the prosodic contrast.

Prosodic feature	**Token A**	**Token B**
Linguistic focus	Dave likes to STUDY music (early focus)	Dave likes to study MUSIC (late focus)
Phrase boundary	If Barbara gives up, the ship will be plundered (early closure)	If Barbara gives up the ship, it will be plundered (late closure)
Lexical stress	COMpound (1st syllable stress)	comPOUND (2nd syllable stress)

Musical beats stimuli were sequences of 18 tones, which consisted of six tones repeated three times. These tones were four-harmonic complex tones with equal amplitude across harmonics and 15-ms on/off cosine ramps. The pitch and duration varied across four levels indicating either a three-note grouping ("strong—weak—weak" pattern, waltz time) or a two-note grouping ("strong—weak" pattern, march time). The strength of these groupings was determined by the increased pitch or duration of the first tone relative to the other tones of the two- or three-note grouping.

Stimuli sampled a 4-by-4 acoustic space across duration and F0 so that the acoustic properties of stimuli cued the appropriate categories (i.e., emphasis on STUDY or MUSIC, early or late phrase closure, lexical stress on first vs second syllable, and musical beats of "strong—weak" or "strong—weak—weak " patterns) to 4 different degrees: 0%, 33%, 67% and 100% (resulting in 4 x 4 grids presented in Figure 3), where 0% values indicate that the F0 or duration characteristics came from Token A recording, 100% means that F0 and duration were identical to the Token B recording, and intermediate values reflect F0 and duration patterns linearly interpolated between the two original recordings. Unlike earlier studies (5x5 grid, Jasmin, Sun, & Tierney, 2021; 7x7 grid, Jasmin, Tierney, Obasih, & Holt, 2022), I do not include the mid-value of ambiguous 50% samples to push participants' responses toward the extreme ends of the continuum[6].

[6] Another practical reason for doing that was to reduce the time needed to complete the task because the experiment was already very long.

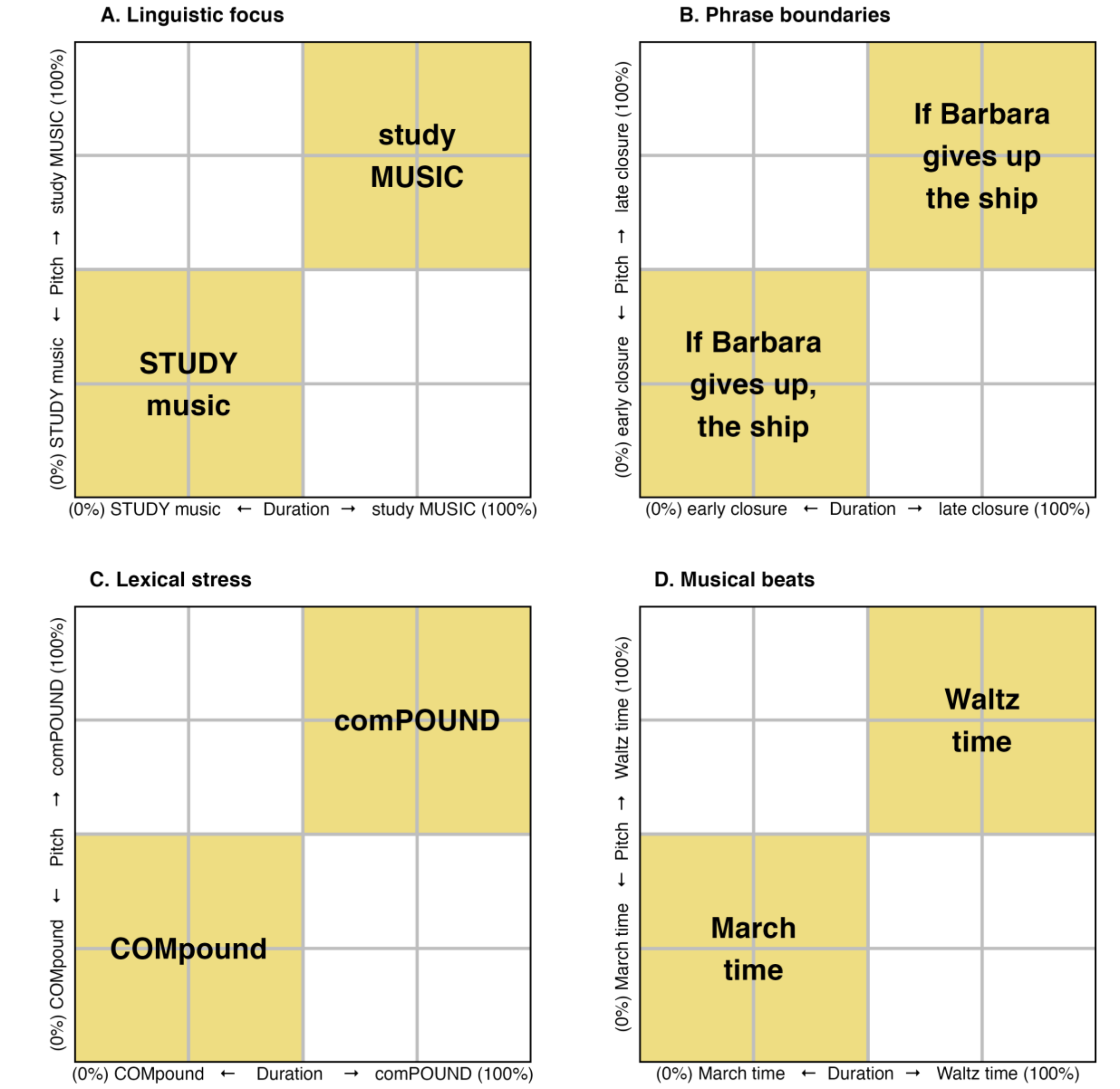

A. Linguistic focus
study MUSIC
STUDY music
(0%) STUDY music ← Pitch → study MUSIC (100%)
(0%) STUDY music ← Duration → study MUSIC (100%)
B. Phrase boundaries
If Barbara gives up the ship
If Barbara gives up, the ship
(0%) early closure ← Pitch → late closure (100%)
(0%) early closure ← Duration → late closure (100%)
C. Lexical stress
comPOUND
COMpound
(0%) COMpound ← Pitch → comPOUND (100%)
(0%) COMpound ← Duration → comPOUND (100%)
D. Musical beats
Waltz time
March time
(0%) March time ← Pitch → Waltz time (100%)
(0%) March time ← Duration → Waltz time (100%)

Figure 3. Cue-weighting task stimulus grids. Participants categorized all the stimuli sampled from a 4-by-4 acoustic space cross duration and pitch (F0). The stimulus space was defined by orthogonal acoustic manipulations across duration and pitch over English speech samples of prosodic feature contrasts (linguistic focus, phrase boundary, and lexical stress) and musical beats.

Procedure

In all four categorization tasks, participants were presented with stimuli that varied orthogonally in the extent to which fundamental frequency (F0, correlate of voice pitch) and duration were indicators of one of the two possible linguistic interpretations. After listening to each stimulus, participants were asked to categorize the stimuli as belonging to one of two categories: phrase with early or late closure ("If Barbara gives up, the ship" vs "If Barbara gives up the ship"), emphasis on the first or second word ("STUDY music" vs "study MUSIC"), lexical stress on the first vs second syllable ("COM-pound" vs "com-POUND"), and musical beats occurring either every two or three notes ("strong—weak" vs "strong—weak—weak" patterns). Before the main task, participants listened to examples of each recording with unaltered pitch and duration and two practice trials with written feedback. The main tasks were identical to the practice except that feedback was no longer provided and all 16 stimuli were presented in random order. There were 10 blocks of each categorization task, which were interleaved in the following order: musical beats, linguistic focus, lexical stress, and phrase boundary. Practice trials were included on the first block of each task but not thereafter. Participants received progress updates after completing one block of each task.

Analyses

I calculated pitch and duration weights by estimating a Firth's biased-reduced logistic regression for each subject (Firth, 1993), with pitch and duration levels (1 through 4) predicting the binary response during each categorization task. The coefficients for pitch and duration were then combined by normalizing them such that they summed to one (Holt & Lotto, 2006; Idemaru et al., 2012; Jasmin et al., 2020), resulting in a normalized perceptual weight between 0 and 1, with values closer to 1 indicating greater reliance on pitch than duration, values closer to 0 indicating the reverse, and 0.5 indicating equal reliance on both features. Analysis was conducted with the implementation of Firth's regression in the logistif R package (Heinze et al., 2022).

Neural measures

EEG data acquisition

EEG data was recorded from 32 Ag-Cl active electrodes using a Biosemi™ ActiveTwo system with the 10/20 electrode montage. Data were recorded at a sampling rate of 16,384 Hz and digitized with a 24-bit resolution. Two external reference electrodes were placed on both earlobes for off-line re-referencing. Impedance was kept below 20 kΩ throughout the testing session. Triggers marking the beginning of each trial (every 6 tones, or 1.2 seconds) were recorded from trigger pulses and sent to the data collection computer. All EEG data processing and analysis were carried out in MATLAB (MathWorks, Inc) using the FieldTrip M/EEG analysis toolbox (Oostenveld et al., 2011) in combination with in-house scripts.

Frequency tagging paradigm

I used the EEG frequency tagging paradigm as a neural measure of dimensional salience. To establish which of the presented dimensions (pitch vs duration) is more salient to participants while listening to verbal and non-verbal sequences, changes in each dimension were tagged to different presentation rates (2.5 or 1.67 Hz). A stronger response to any given frequency represents the salience of a given dimension changing at that rate.

Stimuli

The base stimuli used for the dimensional salience task were the same as those used in the dimension-selective attention task (Table 2). Using the 2 (pitch) x 2 (duration) stimulus grids for each domain, I created 5-Hz sequences (i.e., tone played every 200 ms; 96 seconds in duration) in which tone pitch and duration changed at fixed rates (every two tones, 2.5 Hz, or every three tones, 1.67 Hz). The stimuli consistently varied at these rates with the exception of 20 repetitions which were inserted into each sequence. These repetitions were inserted to prevent the stimuli from becoming overly predictable but were not relevant to the participants' task. For each sequence, the amplitude of 3-5 randomly selected stimuli (32 in total) was decreased by 25% (-12.04 dB) to create amplitude oddballs. Oddball timing was randomized in each sequence, with the exception that oddballs could not occur in the first or last 4.8 seconds (4 epochs) of the sequence and could not occur within 4.8 seconds of another oddball. The same sequences were presented to all participants, but with the order counterbalanced across participants. Stimuli were presented diotically at max 80 dB SPL[7] at a sampling rate of 44100 Hz using PsychoPy3 (v 3.2.3) via insert earphones ER-3A (Etymotic Research, Elk Grove Village IL).

Behavioural task

Participants were asked to listen to these sequences and respond with keyboard presses to occasional quiet tones. The purpose of the behavioural task was to keep participants engaged in listening to the stimuli throughout the session, but without directing their attention to any of the acoustic dimensions of interest.

Before the main task, participants completed a short practice run to familiarize themselves with the task before entering the EEG recording booth. They listened to sequences of verbal and non-verbal sounds for about a minute each and continued until they reached at least 5 out of 6 correct responses without making too many errors to move to the main task. For the practice, the feedback was displayed on the screen, indicating the number of correct and incorrect responses and missed targets. Most participants completed the practice upon their first attempt, and the remaining participants were asked to repeat the practice block to guarantee their good performance during the EEG recording. The main task was identical to the practice but with longer sequences and no visual feedback. Behavioural performance was measured to ensure that participants stayed focused throughout the task. There were 4 blocks, each containing 4 two-minute sequences of sounds.

Behavioral data was computed by calculating the proportion of hits and false alarms and converting them to d-prime, using the loglinear approach to prevent infinite scores (Hautus, 1995). Hits were responses within 1.25 seconds following an oddball, while false alarms were responses outside that time frame divided by the total number of non-oddball tones. Behavioural performance was comparable in both conditions, speech (median d-prime=3.87) and tones (median d-prime=3.67).

[7] Before running the experiment, the peak SPL of each stimulus was set to 80 dB SPL (+/- 1 dB) by scaling the peak amplitude with a custom MATLAB script and checking with the sound level meter. Before each session with participant, I did not test the peak SPL of all stimuli, but only one stimulus per condition.

Intertrial phase coherence (ITPC)

The data were down sampled to 512 Hz and re-referenced to the average of the earlobe reference electrodes. A low-pass zero-phase sixth-order Butterworth filter with a cutoff of 30 Hz was applied. A high-pass fourth-order zero-phase Butterworth filter with a cut-off of 0.5 Hz was then applied and the data epoched (1.2-seconds) based on the recorded trigger pulses. Independent component analysis (ICA) was conducted to correct for eye blinks and horizontal eye movements. Components corresponding to eye blinks and movements were identified and removed based on visual inspection of the time courses and topographies.

Prior to data analysis, I extracted data from the 9 channels with the maximum signal of interest (i.e., ITPC) when averaged across the two rates of dimensional change (pitch at 1.67 Hz/duration at 2.5 Hz and pitch at 2.5 Hz/ duration at 1.67 Hz) and all participants (N=114). The number of channels was decided prior to analysis, following the standard pre-processing procedures. This resulted in a cluster of frontocentral channels (AF4, F3, Fz, F4, FC1, FC2, FC5, Cz, C3) and the data were then averaged across the selected channels. Any remaining artefacts exceeding +/- 100 μV were rejected. Inter-trial phase coherence (ITPC) at the frequencies of dimension change was computed as a measure of cortical tracking of acoustic dimensions. A Hanning-windowed fast Fourier transform was applied to each 1.2-second epoch. The complex vector at each frequency was converted to a unit vector and averaged across trials. The length of the average vector was computed to provide a measure of phase consistency, which ranges from 0 (no phase consistency) to 1 (perfect consistency in phase across trials). The degree of ITPC at the frequency tagged to a certain dimension and EEG signal amplitude provides indices of dimensional salience (i.e., cortical tracking of acoustic dimensions).

General procedure

Participants who expressed interest in the study and responded to the adverts were invited to a short telephone or video call to ensure that they met all the study criteria. Each interview was scheduled individually and during the call, the researcher asked a list of questions about participants' basic demographics, language, and musical background, explained the experimental procedure and task instructions, and answered participants' questions about the study and its purpose. Next, informed consent was obtained from eligible participants, and they received links to two sets of online tasks to complete via the Gorilla Experiment Builder platform (Anwyl-Irvine et al., 2020). Part 1 of the online tasks included a detailed demographics questionnaire, discrimination tasks and dimension-selective attention tasks. Part 2 included the series of categorization tasks. After completing Part 1 and 2, participants were invited to the lab to record their brain activity with EEG[8]. The data analyzed in this chapter corresponds to Time I in Figure 1 (i.e., behavioural and neural data of native English and Mandarin speakers at Pre-Test). Data collection was conducted at the Department of Psychological Sciences at Birkbeck, University of London and all the ethics procedures were approved by the departmental Ethics Committee. All participants were reimbursed for their time in cash (at £10 per hour) or its equivalent in course credits.

Statistical analyses

All statistical analyses were conducted in R. Package lmer4 was used for most mixed-effects regression models (Bates et al., 2015) and glmmTMB (Brooks et al., 2003) for mixed-effects regression models with beta distribution (parameterization of Ferrari & Cribari-Neto, 2004 and betareg package; Cribari-Neto & Zaileis, 2010). Using linear models for continuous outcomes bound by 0-1 intervals might result in spurious effects, so I used a regression model with beta distribution for modeling neural phase locking data. Since discrimination thresholds were not normally distributed, the WRS package was used for a robust variant of ANOVA based on trimmed means (Wilcox & Schonbrodt, 2017). The methods introduced by Wilcox (2017) deal with skewed distributions even when the data is untransformed and when the compared distributions vary in the degree of departure from normality. Multiple comparisons were corrected with False Discovery Rate correction (FDR; Benjamini

[8] Mandarin speakers completed additional behavioural and EEG tasks, several days of online training, post-training assessments and a second EEG session. The data collected during these stages will be discussed in detail in the subsequent chapters of this thesis.

& Hochberg, 1995) from the R stats package. Processed data and analysis scripts can be found at: https://osf.io/5j8pe/?view_only=874b45bef48c4839807867f12a02809a .

2.3 Results

Effects of L1 experience and music training on auditory discrimination thresholds

Visual inspection of plots presenting the untransformed sensitivity thresholds demonstrated an existing trend towards better pitch discrimination in Mandarin non-musicians and comparable thresholds between English and Mandarin musicians (Figure 4).

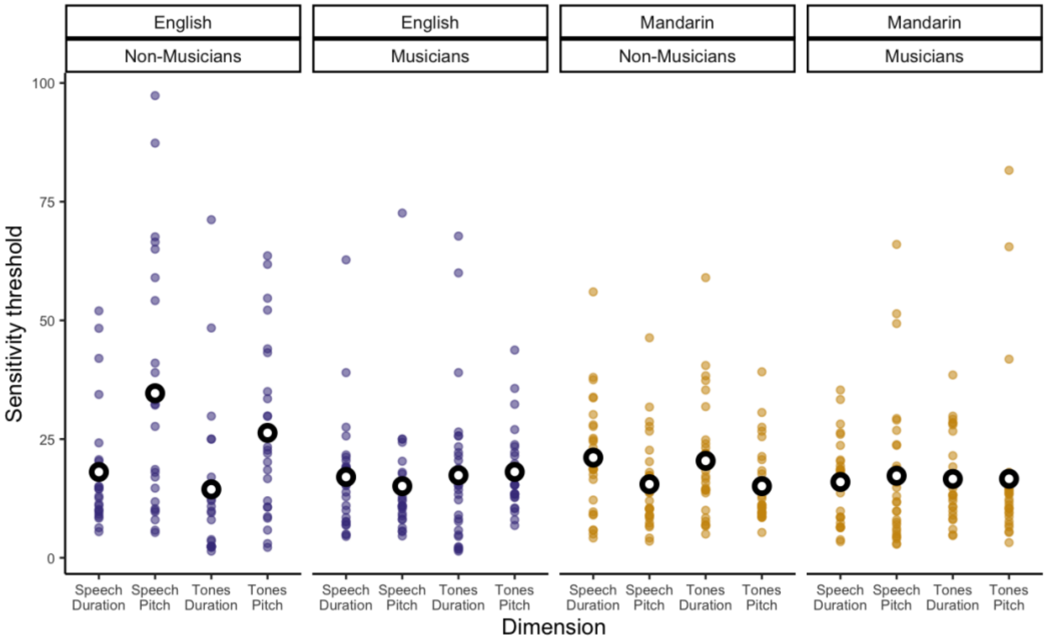


Figure 4. Comparison of raw pitch and duration discrimination thresholds of native English and Mandarin musicians and non-musicians for verbal and non-verbal stimuli.

To statistically compare the differences in sensitivity thresholds to pitch and duration changes between Mandarin and English native speakers with and without musical training, I computed the relative inverse variance in representation of the two dimensions (Ernst & Banks, 2002). Log-transformed thresholds were treated as estimates of perception, and then the variance of the pitch-duration estimates for speech and tones stimuli were computed using the following formula:

$$\hat{S} = \frac{1/(\log(pitch\ threshold)^2)}{1/(log(duration\ threshold)^2)}$$

and used for statistical analyses. A three-way mixed effects robust ANOVA with L1 (Mandarin, English) and music experience (musicians, non-musicians) as between-group factors and stimuli domain (speech, tones) as within-group factor was used to examine the interaction effects between language background, musical expertise and stimulus type on relative perception. Analysis revealed a significant main effect of L1 (F=18.87, p<.001) indicating that the Mandarin participants showed greater relative precision for pitch compared to the English participants across both domains. No main effect of musical expertise or stimulus domain on relative perception estimates (p>.05) was found. A two-way interaction between L1 and musicianship missed significance threshold (F=3.54, p=.06).

To test whether there is a relationship between participants' relative perception of pitch vs duration in speech and non-speech stimuli, I computed Pearson's correlation coefficients for musicians and non-musicians in both language groups. No significant correlations across domains were found ($r_{\text{English musicians}}$=.061, p=.75; $r_{\text{English non-musicians}}$=-.092, p=.66; $r_{\text{Mandarin musicians}}$=.13, p=.49; $r_{\text{Mandarin non-musicians}}$=-.14, p=.46).

Effects of L1 experience and music training on dimension-selective attention

Logistic regression model

The data from the dimension-selective attention tasks were analyzed with a mixed-effects regression model using the glmmTMB function from the glmmTMB package (Brooks et al., 2003). I used a beta distribution that best represented the format of attention data (proportions of correct responses take values from 0 to 1). The categorical variables representing participants' L1 background (English, Mandarin), musicianship (non-musicians, musicians), domain (speech, tones) and attended dimension (duration, pitch) were treatment coded with the first variable level serving as a baseline and the second as a group comparison (0 and 1 respectively). Participants' unique IDs were included as a random intercept. Including random slopes for domain and dimension resulted in overfitting, so a simpler model was used for interpretation.

Results of the mixed-effects regression (Table 4 and Figure 5) suggest that participants' accuracy on the dimension-selective attention task differs depending on their language background (β=-1.06, p=.001), with overall slightly lower performance accuracy among Mandarin speakers. There was also an effect of stimulus domain (β=-1.04, p<.001; better performance for speech stimuli), attended dimension (β=-1.05, p<.001; better performance for pitch) and the interaction between the two (β=1.97, p<.001; better performance accuracy for pitch relative to duration in non-verbal relative to verbal stimuli) hinting at overall differences in attention accuracy across conditions. A significant two-way interaction between L1 and attended dimension (β=2.28, p<.001) suggest superior attention to pitch relative to duration in Mandarin speakers. An interaction between musical training and attended dimension (β=1.86, p<.001) indicates better performance accuracy for pitch relative to duration in musicians compared to non-musicians.

A significant three-way interaction between L1, musical training, and attended dimension (β=-2.40, p<.001) shows that attention differs between musicians vs non-musicians of different L1s, and depends on whether pitch or duration is being measured. A significant interaction between musicianship, stimulus domain and attended dimension (β=-1.28, p=.018) suggests differences amongst musicians in attention to speech and tones depending on which dimension they pay attention to. To interpret these interactions, I ran two regression models for each dimension and L1 and musicianship as predictors. I followed up the significant interaction that emerged for attention to pitch between L1 and musicianship (β=-1.12, p=.002) by reducing the models further (separate model for each L1) and discovered that although there was no difference between musicians and non-musicians among Mandarin speakers (p>.05), English musicians showed a significant advantage over non-musicians (β=1.38, p<.001).

Table 4. Summary of effects in mixed-effects regression model for dimension-selective attention task.

Predictor	***Estimate***	***SE***	***z***	***p***
Intercept	1.650	.251	6.559	**<.001**
L1 (English)	-1.060	.323	-3.262	**.001**
Music (Non-musicians)	-0.064	.336	-.189	.850
Domain (Speech)	-1.043	.282	-3.701	**<.001**
Dimension (Duration)	-1.048	.279	-3.758	**<.001**
L1 x Music	.817	.455	1.797	.072
L1 x Domain	.316	.362	.875	.382
Music x Domain	.553	.378	1.462	.144
L1 x Dimension	2.276	.380	5.993	**<.001**
Music x Dimension	1.856	.391	4.754	**<.001**
Domain x Dimension	1.966	.392	5.019	**<.001**
L1 x Music x Domain	-.114	.507	-.225	.822
L1 x Music x Dimension	-2.399	.534	-4.492	**<.001**
L1 x Domain x Dimension	-.704	.523	-1.346	.178
Music x Domain x Dimension	-1.277	.538	-2.372	**.018**
L1 x Music x Domain x Dimension	.841	.741	1.136	.256

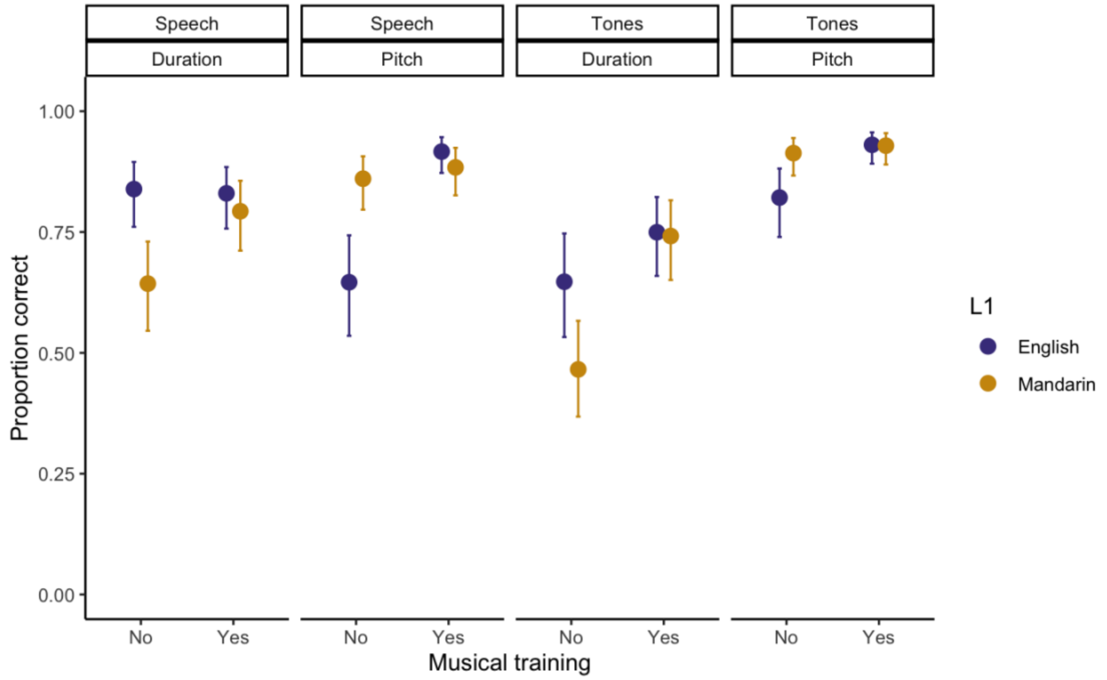


Figure 5. Predicted proportion of correct responses on dimension-selective attention task for Mandarin and English musicians and non-musicians. Responses averaged across participants; error bars – 95%CI.

Correlations across domains

To compare dimension-selective attention across domains (speech, tones) and dimensions (pitch, duration), I computed normalized pitch accuracy by dividing attention to pitch performance by the sum of attention to pitch and attention to duration performance and used these values for correlation analysis. Since the variables were not normally distributed as established by Shapiro-Wilk test ($p<.05$), I used Spearman's correlation. Normalized pitch accuracy was positively correlated across domains for English musicians ($\rho=0.65$, $p<.001$) and Mandarin non-musicians ($\rho=0.38$, $p<.036$). Visual inspection of the data indicates that English speakers' normalized accuracy is more clustered around 0.5 possibly reflecting more balanced and accurate performance across conditions than Mandarin speakers (Figure 6).

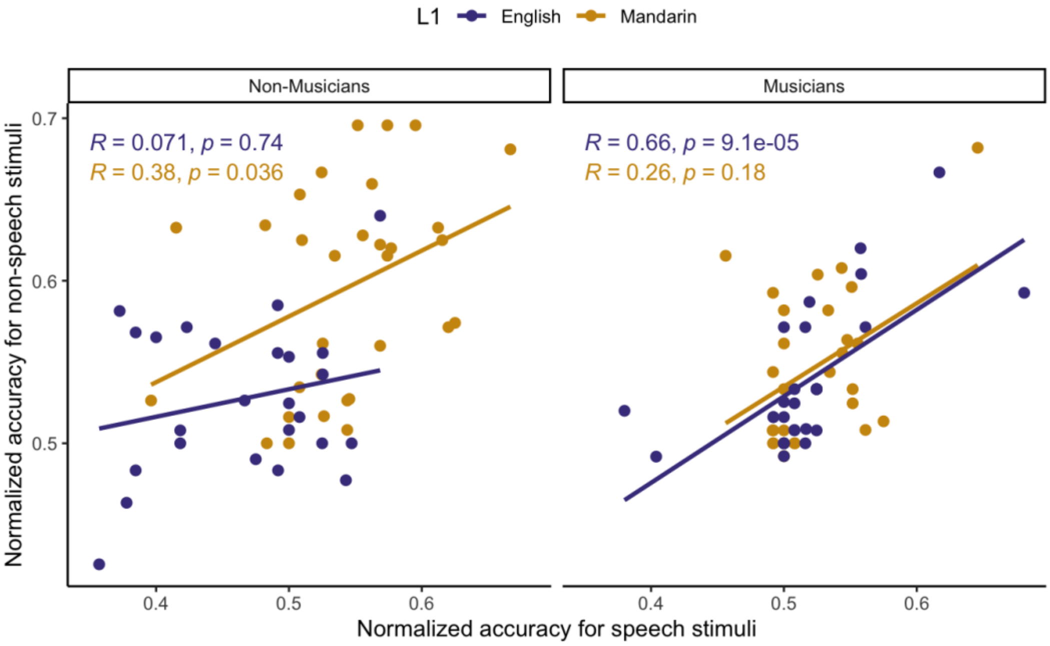


Figure 6. Performance on dimension-selective attention task for musicians and non-musicians. Correlation between normalized accuracy for speech and tones stimuli for L1 Mandarin and English musicians and non-musicians.

Effects of L1 experience and music training on cue weighting strategies

Differences in normalized cue weights

Normalized cue weights differed between the native speakers of Mandarin and English. Visual inspection of the data suggests that Mandarin speakers had larger normalized pitch cue weights than English speakers for all linguistic features (Figure 7).

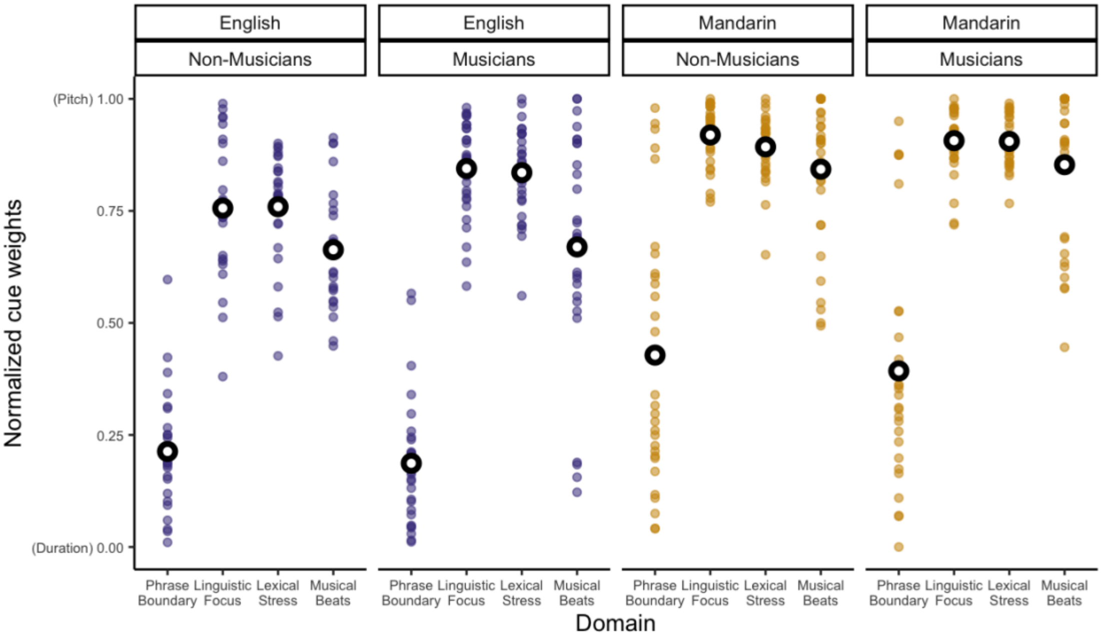


Figure 7. Normalized cue weights across domains. Compared to musicians, English non-musicians rely less on pitch in categorizing linguistic focus and lexical stress, whereas Mandarin speakers' performance is comparable. Response patterns observed in English musicians are similar to those of Mandarin speakers (musicians or non-musicians).

Logistic regression models

To quantify listeners' use of acoustic cues in categorization, the trial-by-trial categorization data were analyzed with a series of mixed effects logistic regression models using the glmer function from the lme4 package (Bates et al., 2015). The dependent variable was the response on each trial (0–incorrect, 1–correct). The categorical variables representing participants' L1 background (English, Mandarin) and musical training (non-musicians, musicians) were treatment coded with the first variable level serving as a baseline and the second as a group comparison (0 and 1 respectively). The continuous predictors pitch level (1-4) and duration level (1-4) were standardized by centering and dividing by 2 standard deviations using the rescale function from the arm R package (Gelman et al., 2021). The resulting beta coefficients from the model represent the change in log odds given an increase of one standard deviation of that variable. Participants' unique IDs were included as a random intercept[9]. Results of the mixed-effects logistic regression models are presented in Table 5.

Table 5. Summary of effects in mixed effects logistic regression models for categorization tasks

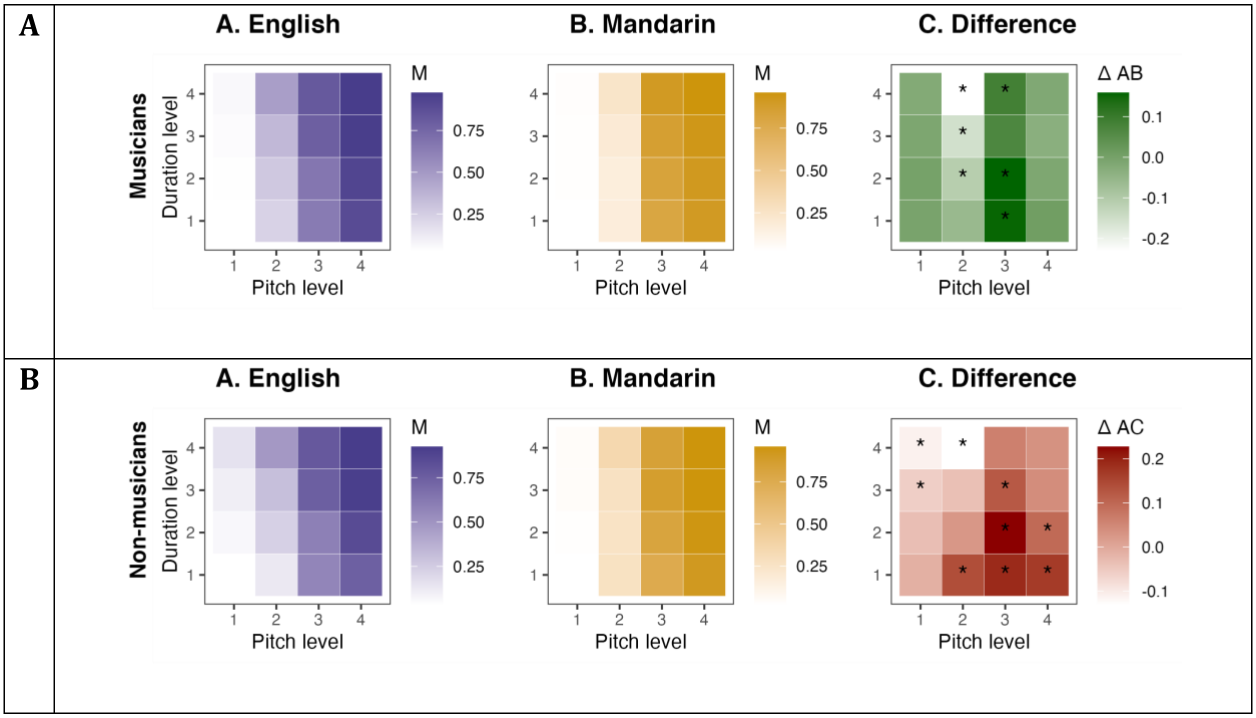
Task	***Predictor***	***Estimate***	***SE***	***z***	***p***
Linguistic Focus					
	Intercept	-.165	.167	-.990	.322
	L1 (English)	.401	.225	1.782	.075
	Music (non-musicians)	.313	.228	1.369	.171
	Pitch	3.618	.110	32.889	**<.001**
	Duration	1.174	.089	13.184	**<.001**
	L1 x Music	-.508	.316	-1.610	.107
	L1 x Pitch	1.958	.185	10.567	**<.001**
	Music x Pitch	1.427	.177	8.053	**<.001**
	L1 x Duration	-.707	.128	-5.530	**<.001**
	Music x Duration	-.216	.129	-1.677	.093
	Pitch x Duration	-.113	.208	-.545	.586
	L1 x Music x Pitch	-1.384	.279	-4.967	**<.001**
	L1 x Music x Duration	.211	.184	1.148	.251
	L1 x Pitch x Duration	.466	.339	1.373	.170
	Music x Pitch x Duration	.540	.322	1.676	.094
	L1 x Music x Pitch x Duration	-.288	.511	-.564	.573
Phrase Boundary					
	Intercept	-.224	.098	-2.285	**.022**
	L1 (English)	-.090	.130	-.694	.488
	Music (non-musicians)	-.320	.136	-2.351	**.019**
	Pitch	1.176	.094	12.520	**<.001**
	Duration	4.176	.123	34.007	**<.001**
	L1 x Music	.168	.184	.911	.362
	L1 x Pitch	.512	.120	4.272	**<.001**

[9] Including random slopes for pitch level and duration level and their interaction results in overfitting, so the simpler model without random slopes was selected across categorization tasks.

Music x Pitch	.028	.137	.203	.839
L1 x Duration	-1.920	.145	-13.228	**<.001**
Music x Duration	1.045	.189	5.538	**<.001**
Pitch x Duration	-.059	.236	-.252	.801
L1 x Music x Pitch	.186	.178	1.047	.295
L1 x Music x Duration	-.329	.224	-1.469	.142
L1 x Pitch x Duration	.105	.282	.371	.711
Music x Pitch x Duration	-.109	.361	-.303	.762
L1 x Music x Pitch x Duration	-.083	.432	-.193	.847
Lexical Stress				
Intercept	-.005	.170	-.028	.978
L1 (English)	-.210	.230	-.914	.361
Music (non-musicians)	-.101	.233	-.434	.664
Pitch	2.980	.096	30.928	**<.001**
Duration	.974	.082	11.842	**<.001**
L1 x Music	.003	.323	.010	.992
L1 x Pitch	2.491	.174	14.307	**<.001**
Music x Pitch	1.597	.160	9.954	**<.001**
L1 x Duration	-.407	.123	-3.311	**<.001**
Music x Duration	-.052	.121	-.428	.669
Pitch x Duration	-.030	.183	-.165	.869
L1 x Music x Pitch	-.860	.278	-3.101	**<.001**
L1 x Music x Duration	-.076	.182	-.418	.676
L1 x Pitch x Duration	.187	.328	.572	.568
Music x Pitch x Duration	-.167	.284	-.590	.555
L1 x Music x Pitch x Duration	.895	.512	1.747	.081
Musical Beats				
Intercept	-.424	.202	-2.098	**.036**
L1 (English)	.173	.274	.632	.528
Music (non-musicians)	.032	.275	.115	.908
Pitch	6.442	.221	29.165	**<.001**
Duration	2.656	.133	19.919	**<.001**
L1 x Music	.010	.382	.027	.979
L1 x Pitch	3.396	.386	8.791	**<.001**
Music x Pitch	-1.094	.272	-4.029	**<.001**
L1 x Duration	-.825	.193	-4.274	**<.001**
Music x Duration	-.037	.176	-.210	.834
Pitch x Duration	5.025	.377	13.347	**<.001**
L1 x Music x Pitch	.034	.501	.068	.946
L1 x Music x Duration	-.112	.261	-.428	.667
L1 x Pitch x Duration	.537	.620	.866	.387
Music x Pitch x Duration	-2.470	.467	-5.269	**<.001**
L1 x Music x Pitch x Duration	.576	.826	.698	.485

Linguistic focus

Results from logistic regression (Table 5 and Figure 8C) show that participants' categorization of the linguistic focus stimuli was influenced by both pitch (β=3.62, p<.001) and duration (β=1.17, p<.001). Two-way interaction effects between L1 and both features also suggests that Mandarin speakers relied more on pitch (β=1.96, p<.001) and less on duration (β=-0.71, p<.001) than English speakers when making these decisions. There was no significant interaction between pitch and duration. A significant interaction between musicianship and pitch level suggests greater reliance on pitch by musicians (β=1.43, p<.001), however a three-way interaction between L1, musicianship and pitch level (β=1.38, p<.001) also indicates that Mandarin and English musicians and non-musicians differ in their pitch reliance. To follow up on this interaction, I ran two separate regression models for Mandarin and English speakers. This post hoc analysis revealed that there was no difference between Mandarin musicians and non-musicians in their pitch use (p>.05), but English musicians relied on pitch more than non-musicians (β=1.45, p<.001). Response patterns demonstrate that the differences in responses between Mandarin and English non-musicians were observed predominantly in the corners of the stimulus space where the two cues conflict (Figure 8AB).



C

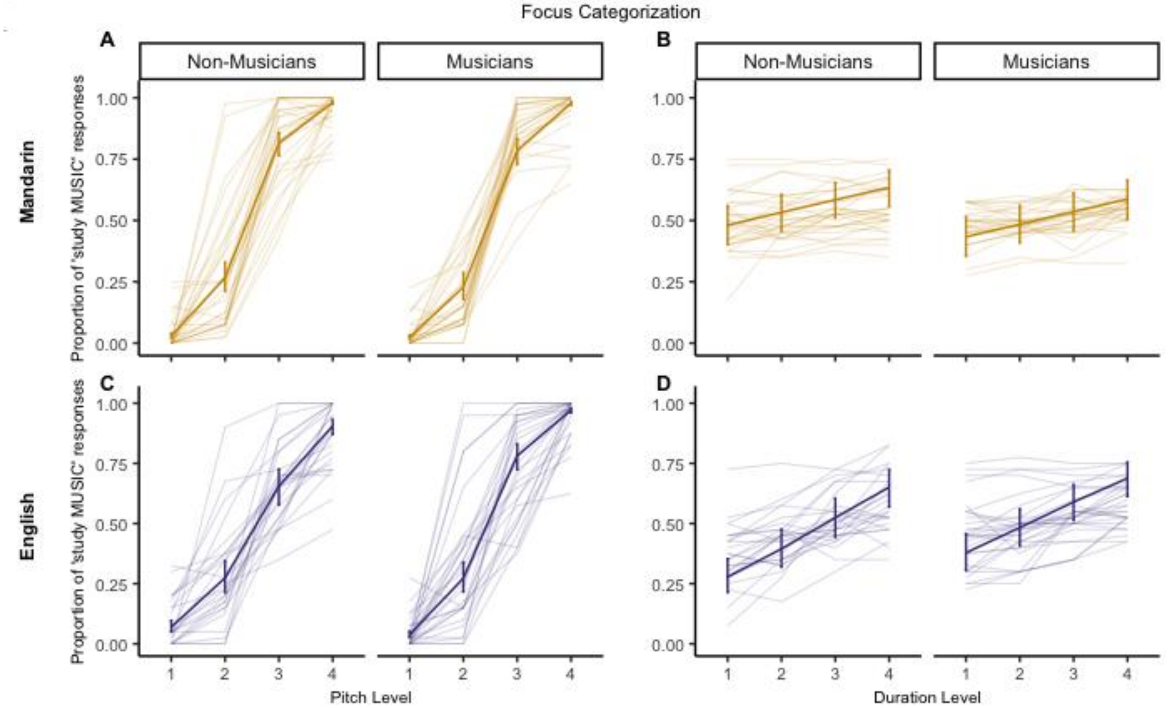


Figure 8. Linguistic focus categorization response patterns. (8AB, ABC) Mean responses pattern of English and Mandarin speakers and their differences plotted separately for musicians and non-musicians. All differences marked with asterisks were significant as shown by Mann-Whitney U tests (FDR-corrected). (8C, AC) Predicted proportion of "study MUSIC" vs "STUDY music" as a function of pitch level for A) Mandarin and C) English speakers. (8C, BD) Predicted proportion of "study MUSIC" vs "STUDY music" as a function of duration level for B) Mandarin and D) English. Dark lines – predicted responses; light lines – individual responses; error bars – 95%CI.

Phrase Boundary

Results from logistic regression (Table 5 and Figure 9C) indicated that during categorization of phrase boundary stimuli participants were influenced by both acoustic features, pitch (β=1.18, p<.001) and duration (β=4.18, p<.001). There was no significant interaction between pitch and duration. Although there was no main effect of L1 group, two-way interaction effects between L1 and both features suggests that Mandarin speakers relied more on pitch (β=0.51, p<.001) and less on duration (β=-1.92, p<.001) than English speakers when making these decisions. The main effect of musicianship indicates differences in categorization driven by musical training (β=-.32, p=.019). A two-way interaction between musicianship and duration conveys that musicians are likely to rely on this feature more than non-musicians (β=1.05, p<.001). The difference between Mandarin and English musicians and non-musicians was observed predominantly in the corners of the stimulus space where the two cues conflict (Figure 9AB).

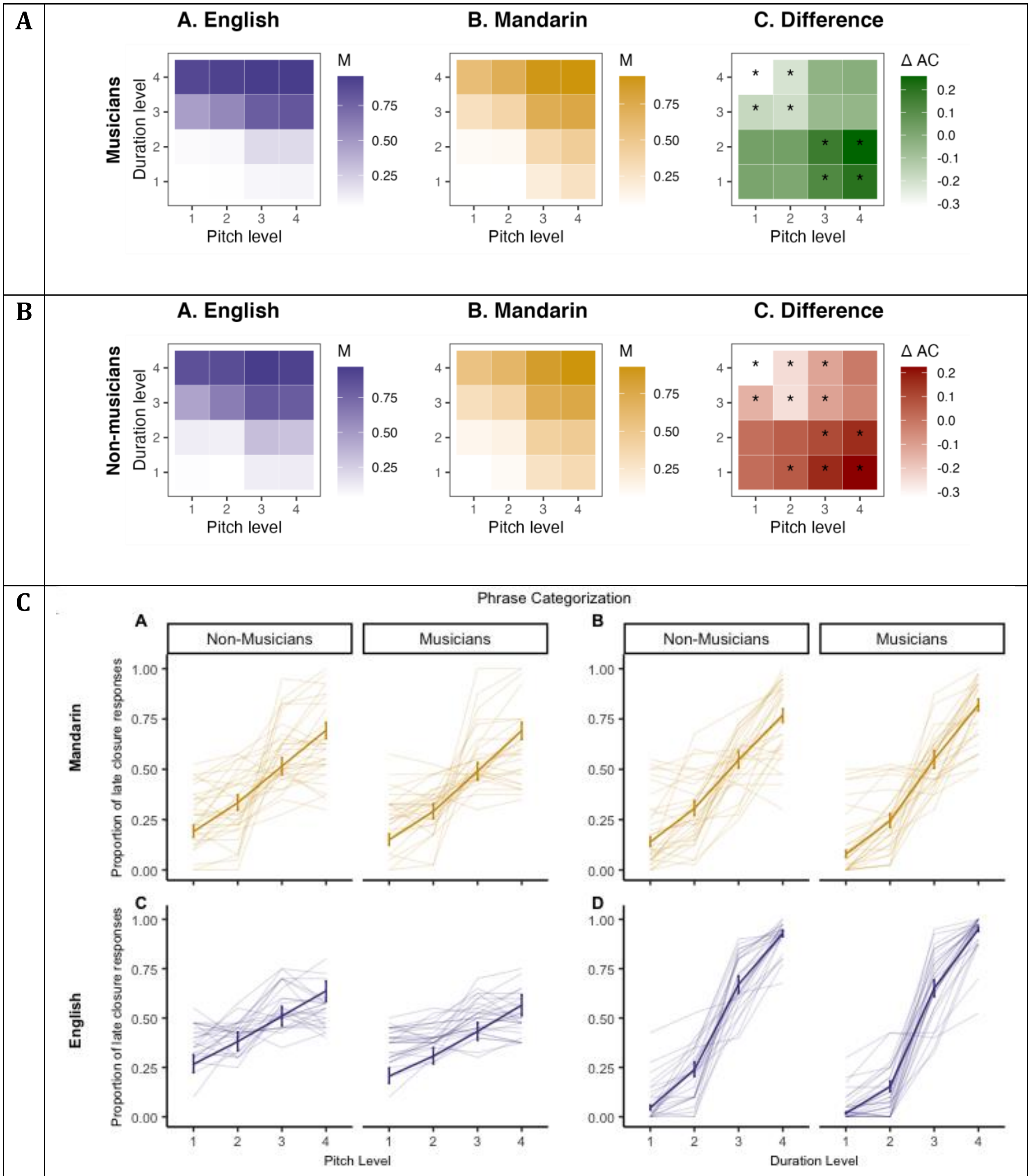


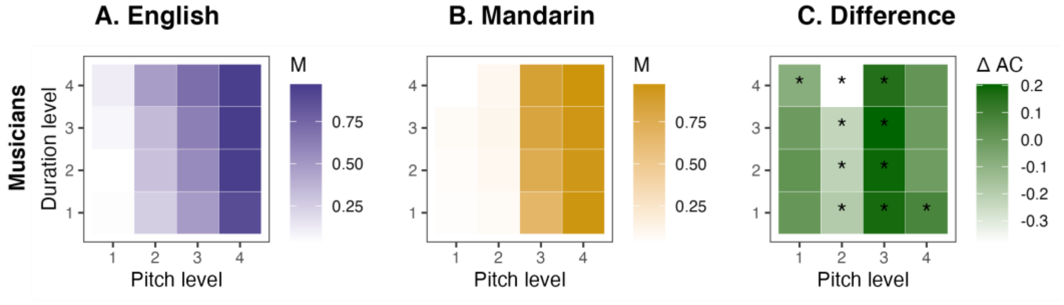
Figure 9. Phrase boundary categorization responses patterns. (9AB, ABC) Mean responses pattern of English and Mandarin speakers and their differences plotted separately for musicians and non-musicians. All differences marked with asterisks were significant as shown by Mann-Whitney U tests (FDR-corrected). (9C, AC)

Predicted proportion of late vs early closure as a function of pitch level for A) Mandarin and C) English. (9C, BD) Predicted proportion late vs early closure as a function of duration level for B) Mandarin and D) English. Dark lines – predicted responses; light lines – individual responses; error bars – 95%CI.

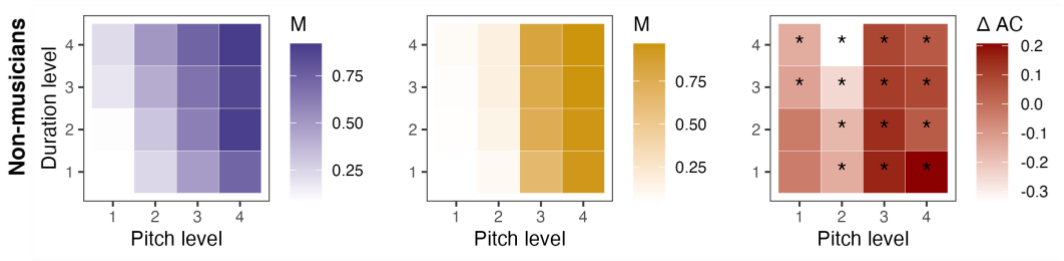
Lexical Stress

Results from logistic regression (Table 5 and Figure 10C) demonstrated that participants' categorization of lexical stress was influenced by both acoustic features pitch (β=2.98, p<.001) and duration (β=0.97, p<.001). The interaction effect between L1 and both features suggests that Mandarin speakers relied more on pitch (β=2.49, p<.001) and less on duration (β=-0.41, p<.001) when making these decisions. A two-way interaction between musicianship and pitch level suggests stronger reliance on pitch by musically trained participants (β=1.60, p<.001). A three-way interaction between L1, musicianship and pitch level (β=-.86, p=.002), suggests differences in pitch reliance by musicians and non-musicians speaking different L1s. To interpret this interaction, I ran post hoc analyses with two regression models, one for each L1 group. Results indicated that when compared to non-musicians, musicians relied more on pitch among Mandarin (β=.72, p=.001) and English speakers (β=1.60, p<.001). The differences in response patterns between Mandarin and English speakers can be seen in the corners of the stimulus space where the two cues conflict (Figure 10AB).

A



B



C

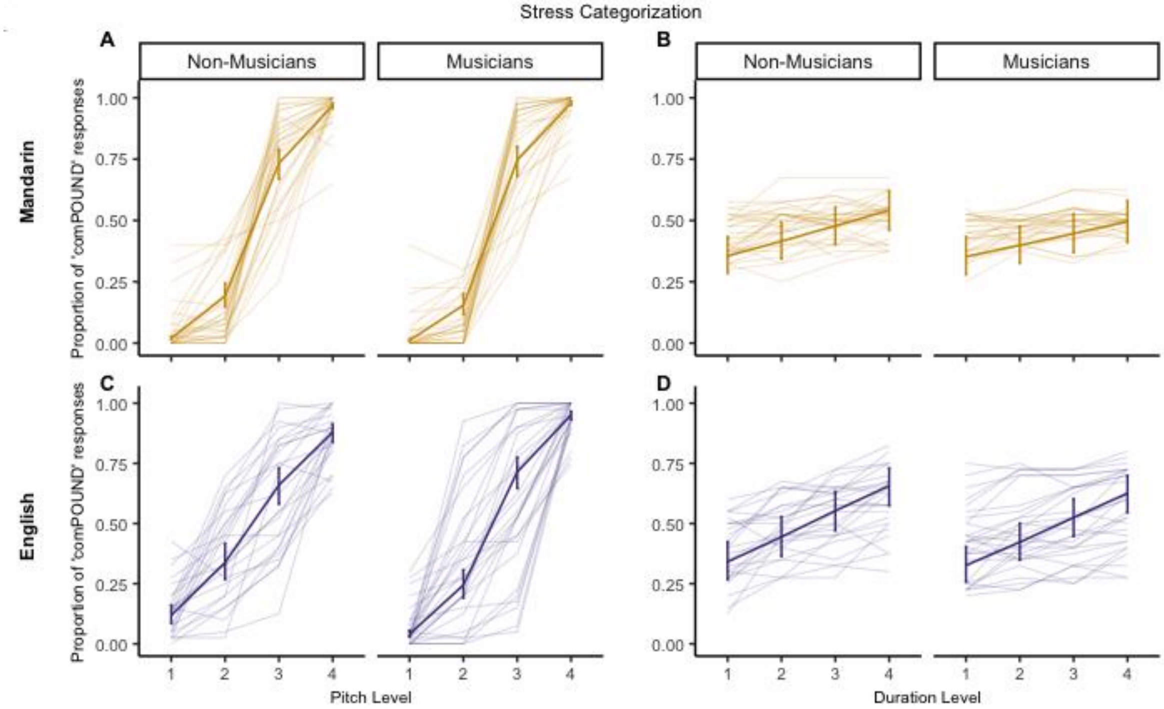
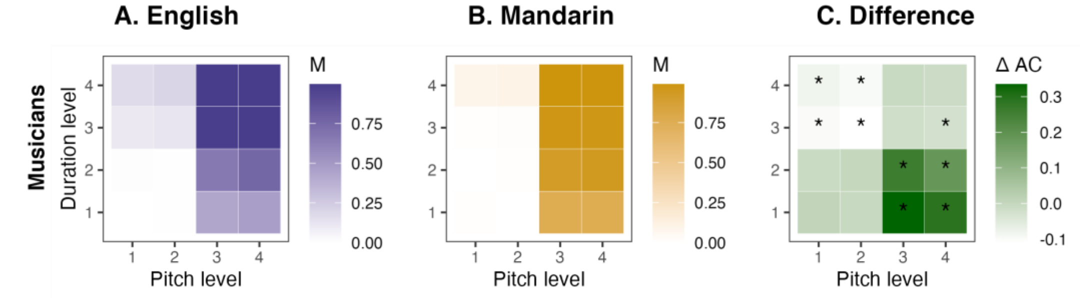


Figure 10. Lexical stress categorization responses patterns. (10AB, ABC) Mean responses pattern of English and Mandarin speakers and their differences plotted separately for musicians and non-musicians. All differences marked with asterisks were significant as shown by Mann-Whitney U tests (FDR-corrected). (10C, AC) Predicted proportion of responses as a function of pitch level for A) Mandarin and C) English. (10C, BD) Predicted proportion of responses as a function of duration level for B) Mandarin and D) English. Dark lines – predicted responses; light lines – individual responses; error bars – 95%CI.

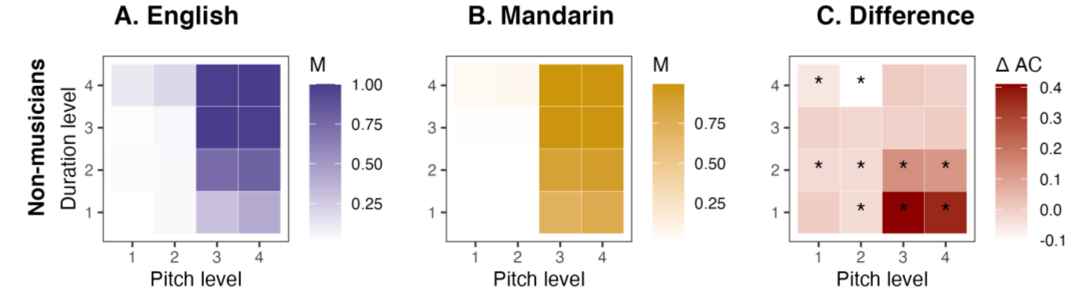
Musical Beats

Results from logistic regression (Table 5 and Figure 11C) show that when categorizing musical beats participants were influenced by both acoustic features, pitch (β=6.44, p<.001) and duration (β=2.66, p<.001). There was also a significant interaction between pitch and duration (β=5.02, p<.001), suggesting the interdependence of both types of information in categorizing musical beats. Two-way interaction effects between L1 and both features suggest that Mandarin speakers relied more on pitch (β=3.40, p<.001) and less on duration (β=-1.09, p<.001) when making these decisions. A two-way interaction between musicianship and pitch level indicates less reliance on pitch by musicians (β=-1.09, p<.001) when compared to non-musicians. A significant three-way interaction between musicianship and interaction between pitch and duration suggests that participants with musical training might rely less on the combination of those features than non-musicians (β=-2.47, p<.001).

A



B



B

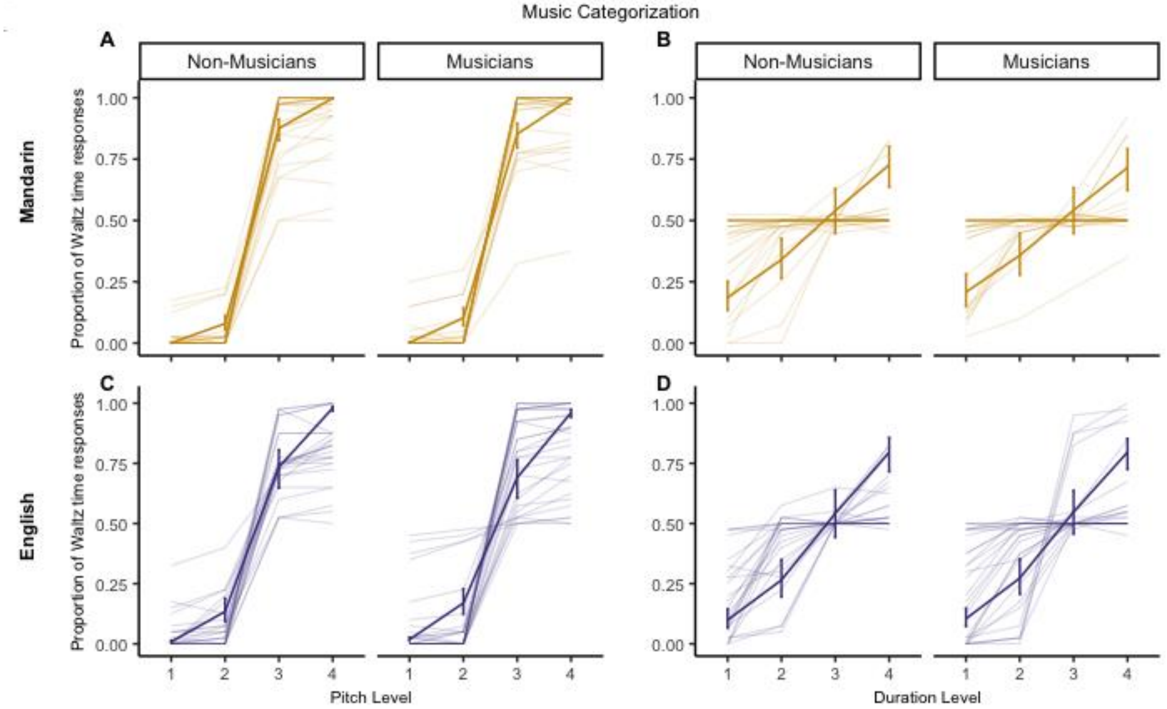


Figure 11. Musical beats categorization responses patterns. (11AB, ABC) Mean responses patterns of English and Mandarin speakers and their differences plotted separately for musicians and non-musicians. All differences marked with asterisks were significant as shown by Mann-Whitney U tests (FDR-corrected). (11C, AC) Predicted proportion of responses as a function of pitch level for A) Mandarin and C) English. (11C, BD) Predicted proportion of responses as a function of duration level for B) Mandarin and D) English. Dark lines – predicted; light lines – individual responses; error bars – 95%CI.

Effects of L1 experience and music training on dimensional salience

To explore the role of language background and musical training in neural tracking of acoustic dimensions, I analyzed the ITPC data with the mixed-effects regression model. I used the glmmTMB function from the glmmTMB R package (Brooks et al., 2003) with a beta distribution appropriate for continuous ITPC data bound by the 0-1 interval. The dependent variable was the mean ITPC. The categorical variables representing participants' L1 background (English, Mandarin), musical training (non-musicians, musicians), domain (speech, tones), and acoustic dimension (duration, pitch) were treatment coded (0 for reference level, 1 for comparison level). Participants' unique IDs were included as a random intercept.

Results from the mixed-effects regression model (Table 6 and Figure 12) demonstrated that participants' phase locking varied depending on domain (weaker tracking of non-verbal stimuli; β=-.49, p<.001) and dimension (weaker pitch tracking; β=-.599, p<.001), as well as their significant interaction (β=.711, p<.001). A significant two-way interaction between L1 and musical training (β=.386, p=.005) emphasizes the role of language background and musical expertise in neural tracking of acoustic dimensions. I followed up on this interaction with post hoc analysis. Reduced regression models for each L1 group with musicianship as a predictor revealed that Mandarin musicians are showing more overall phase-locking compared to the Mandarin non-musicians (β=.21, p=.002), whereas there is no difference between English musicians and non-musicians (p>.05).

Table 6. Summary of effects in mixed effects regression models for ITPC

Predictor	***Estimate***	***SE***	***z***	***p***
Intercept	-1.704	.073	-23.199	**<.001**
L1 (English)	-.085	.100	-.858	.391
Music (Non-musicians)	-.129	.101	-1.268	.205
Domain (Speech)	-.490	.088	-5.573	**<.001**
Dimension (Duration)	-.599	.090	-6.629	**<.001**
L1 x Music	.386	.138	2.786	**.005**
L1 x Domain	.132	.118	1.112	.266
Music x Domain	.072	.122	.589	.556
L1 x Dimension	.039	.123	.317	.751
Music x Dimension	.160	.124	1.292	.196
Domain x Dimension	.711	.130	5.479	**<.001**
L1 x Music x Domain	-.168	.165	-1.023	.306
L1 x Music x Dimension	-.196	.169	-1.161	.245
L1 x Domain x Dimension	-.247	.176	-1.401	.161
Music x Domain x Dimension	-.118	.178	-.662	.508
L1 x Music x Domain x Dimension	.215	.243	.883	.377

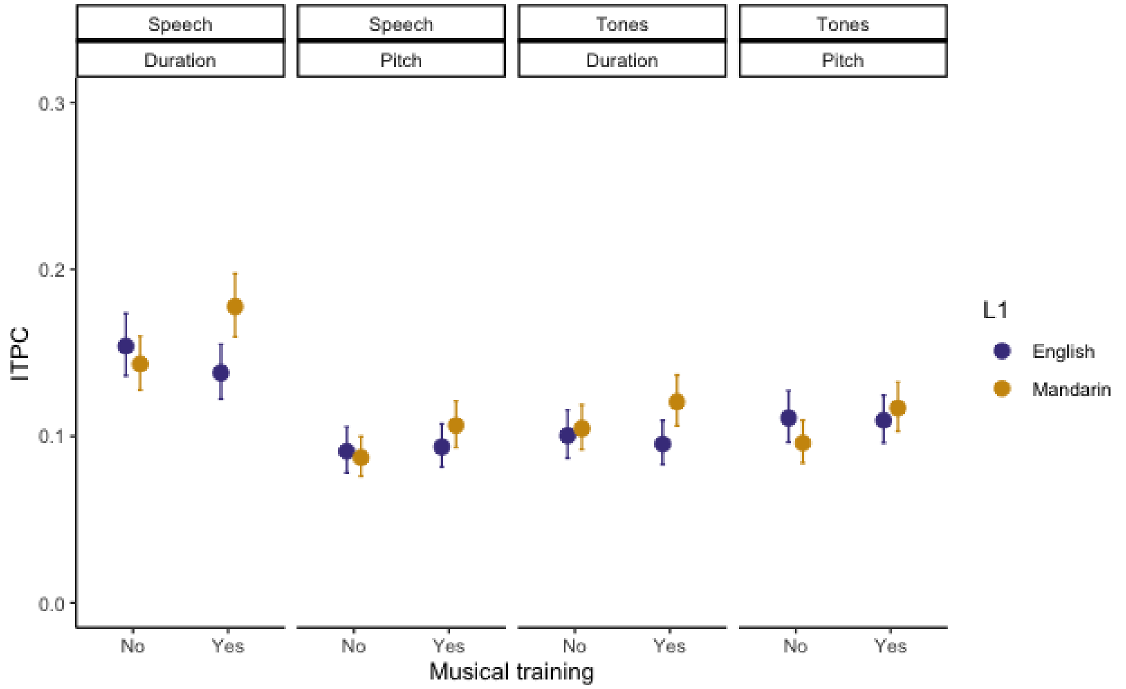


Figure 12. Average ITPC of Mandarin and English musicians and non-musicians at the frequencies corresponding to variations in duration and pitch for each domain (verbal, non-verbal) across the frontocentral channels selected for analysis.

To compare the dimensional tracking across domains (speech, tones), I estimated the measure of relative salience by dividing pitch tracking by the sum of pitch and duration tracking and used these values for correlation analysis. Both variables were normally distributed according to the Shapiro-Wilk test (p>.05), so Pearson's correlation was used. The relative salience of pitch versus duration was positively correlated for English non-musicians (r=.52, p=.007) and musicians (r=.39, p=.037), and Mandarin musicians (r=.40, p=.031), indicating similar salience distribution across domains (Figure 13). The data of Mandarin non-musicians showed a trend towards a positive correlation but did not reach the significance threshold (p=.11).

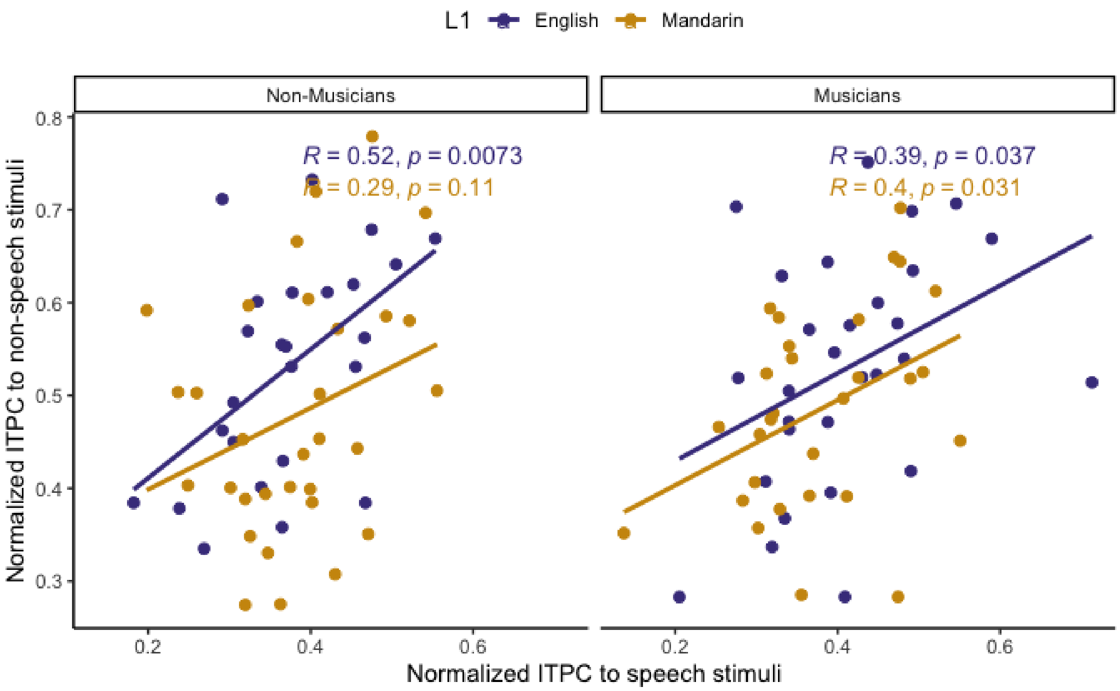


Figure 13. Relative dimensional salience between verbal and non-verbal stimuli for Mandarin and English musicians and non-musicians.

2.4 Discussion

This study investigated the effects of L1 background and musical experience on auditory sensitivity thresholds, dimension-selective attention, cue weighting strategies and dimensional salience. Based on attentional theories of cue weighting (Francis et al., 2000; Francis & Nusbaum, 2002), I predicted that listeners would up-weight the dimension that is especially relevant in their L1 (i.e., pitch for native Mandarin speakers) or was learnt as a part of the training because of its importance for conveying musical structures (i.e., pitch for musicians). Furthermore, I also expected that they would show superior performance in behavioural tasks requiring the use of that dimension and increased salience of that dimension reflected by its more robust cortical tracking. Overall, these findings support my predictions and reveal differences between tonal and non-tonal language speakers' behavioural performance and listening strategies – Mandarin speakers indeed show enhanced pitch processing across a range of behavioural tasks, but not at the neural level. Building on the effects of language background, musical training shows L1-specific effects.

Perceptual strategies depend on language background

Confirming my expectations, I show that Mandarin speakers exhibit consistent pitch bias across speech categorization tasks, whereas English speakers more flexibly adjust between the optimal cues for a given task. These results are consistent with the hypothesis that perceptual strategies change through a lifetime of L1 experience to support smooth native communication and can be suboptimal in the context of new languages (e.g., Jasmin, Sun, Tierney, 2021). Despite the consistent preference for pitch among Mandarin speakers, I also observed a high degree of individual variability in their responses (see Figure 7 for individual differences in normalized cue weights across categorization tasks). For example, some individuals had an extreme pitch bias during phrase boundary categorization or relied less on pitch while categorizing other stimuli where it was the most useful cue (i.e., lexical stress and linguistic focus). These results strongly indicate that besides language and musical expertise, individual factors might contribute to shaping perceptual strategies (e.g., attentional control and working memory, Ou, Law, & Fung, 2015; attentional switching, Ou & Law, 2017).

Although I show that the effects of L1 and musical background are not limited to speech stimuli, I do not offer sufficient evidence to claim that these effects generalize to music processing, as the beat-like characteristics of my stimuli that are structurally more similar to word or syllable emphasis are insufficient to capture the complexity of music. Future studies should properly examine the role of cue weighting in music perception by including more music-like stimuli. For example, contrasts comprising musical cadences or chord transitions that create a sense of full or partial resolution in musical phrases could be more appropriate for approximating cue weighting in melody perception. Cadential segments, similarly to speech, can be described by multiple cues, for example, by slowing down at structural endings and pitch or harmonic movement toward more stable chords (Palmer & Krumhansl, 1987). Cue weighting in rhythm could be estimated using sequences performed on Yoruba talking drums – dundun drums produce rhythms and sounds similar to speech tones by loosening or tightening the tension of the drum membranes (Durojaye et al., 2021). As the dynamics of how acoustic information is integrated in music perception are not well understood, developing a model of cue weighting in music perception would also support that goal.

Musical training benefits differ between English and Mandarin speakers

A comparison of cue weighting strategies between musicians and non-musicians revealed that musical training seems to affect English musicians' weighting in stress and focus perception by sharpening their tuning to primary dimensions, consistent with results presented by Symons and Tierney (2023). They showed that apart from the preferential weighting of more familiar or reliable dimensions, people also differ in the degree to which they rely on the primary dimension, regardless of which dimension that is. Similarly, I found stronger reliance on primary cues for English musicians in focus and stress categorization tasks but not for phrase stimuli. Such a pattern did not emerge for L2 speakers, suggesting that more advanced L2 knowledge might be required to develop more native-like strategies or that Mandarin speakers are already at ceiling on their use of pitch cues in categorization of stress and focus stimuli.

Also the relative perception of pitch and duration as measured by discrimination tasks by musicians and non-musicians differed significantly between English and Mandarin speakers. While English musicians benefited most in pitch perception in speech and tones, Mandarin musicians showed marginal improvements in speech and tone duration discrimination compared to non-musicians, but no effects of musical training were observed for pitch across domains. This is perhaps because Mandarin speakers are already at ceiling in pitch discrimination (e.g., Giuliano et al., 2011), and no further enhancements are possible. My results clearly show that musical training has differential effects on people that depend on their inherent auditory abilities.

It is well-acknowledged that musical training leads to improvements in various aspects of auditory processing (Tervaniemi, 2009). However, the extent and nature of these enhancements might be specific to the type of auditory exposure (Micheyl et al.,2006; Zaltz, Globerson, & Amir, 2017). For example, research showed that both musicians and audio engineers have generally lower pitch sensitivity thresholds than those without any training (Caprini et al., 2021). However, it seems that the patterns of advantage exhibited by listeners are largely bound by the specific characteristics of training they obtained – while musicians and engineers performed similarly in pitch discrimination tasks, they exhibited differences in sustained selective attention and sound memory tasks (Caprini et al., 2021). On the other hand, type of musical training did not have any differential effects on participants' ability to synchronize to a rhythm or beat (Matthews, Thibodeau, Gunther, & Penhune, 2016). Further investigations are needed to determine whether the focus of the targeted training could have differential effects on perceptual strategies. An interesting avenue for future research would be to consider the effects of musical training focused on melodic structure (pitch-based) compared to training concentrated on temporal aspects of music (rhythm-based) or even vocal training (i.e., singing vs voice acting) and their role in shaping listening strategies. Differences arising from different types of training (e.g., self-taught or trained in the specific music education method) or cultural traditions (e.g. Western vs traditional Chinese) should also be examined.

Salience and attention as underlying mechanisms

Finally, I asked whether L1 background and musical training are linked to one's ability to selectively attend to various acoustic dimensions and their enhanced salience. I hypothesized that the extensive pitch expertise of Mandarin speakers would result in an increased salience of this dimension across domains. However, as shown by cortical tracking of pitch vs duration in the current data, this is not necessarily the case – I did not find any dimension-specific effects. This finding undermines the role of dimensional salience as a potential driver of differences in perceptual strategies and leaves the question of what contributes to the observed differences open for discussion.

I found more overall phase-locking among Mandarin musicians compared to non-musicians, but no such differences were present when comparing English musicians and non-musicians. One potential explanation of this finding is that Mandarin speakers are more used to tracking acoustic changes at the syllable level (i.e., changing pitch contour of tones) than English speakers who do not have such an experience. This could also be a result of the specific musical training Mandarin learners obtained in their home countries. A recent comparative study between Chinese traditional and Western music emphasized several important differences (Hao, 2023). For example, Chinese music places more emphasis on melodic structure, whereas Western music focuses more on rhythm and harmony. Previous research also showed that there were no differences between participants with little and without musical training in their cortical responses to music and speech rhythm, but their response to musical rhythm tended to increase with a growing number of years of training (e.g., Harding et al., 2019). In my analysis, I considered musicianship as a categorical variable, but perhaps there were some systematic differences in the amount and quality of musical training between the L1 groups that might be driving this effect.

I also need to remember that speech perception involves the processing of not only acoustic but also linguistic information. It could be that in the absence of appropriate context, the salience of acoustic dimensions does not transpire as relevant or reliable to the listener in any way. One way to test that would be to conduct a follow-up study using an experimental paradigm similar to the one reported in this study, but with linguistically meaningful stimuli (e.g., one syllable words). In the recent study (Daube, Ince, & Gross 2019) the authors argued that acoustic information should be sufficient to address questions about understanding neural responses to speech as they were able to reliably decode subgroups of phonemes from MEEG data using their simplest feature model including speech envelope tracking. However, unlike in this study, they measured brain responses to an hour of continuous speech, not isolated vowels. It is also important to note that the results of their study do not necessarily negate the role of linguistic information in speech encoding, but rather highlight the potential contributions of acoustic information to this process.

Recent advances in neuroimaging data analysis techniques (i.e., multivariate temporal response function [mTRF], Crosse et al., 2016, 2021) allow for disentangling acoustic and linguistic features' individual contributions in predicting the brain response to complex stimuli (e.g., semantic dissimilarity, Broderick et al., 2018; acoustic vs phonetic features, Desai, Field, & Hamilton, 2023). One potential avenue for future research could entail tracking of acoustic dimensions within naturalistic speech and music samples, as opposed to isolated vowels or tones concatenated into meaningless sequences. This approach could involve analyzing various forms of vocalizations, such as native and non-native speech, singing or even speech-to-song transformations, to better understand how different acoustic characteristics are represented by speakers of different languages and encoded in the brain at different time scales and how various types of information contribute to parsing continuous sounds streams into meaningful units, resolving stimulus identity and intended meaning. Such investigations could offer valuable insights into the perceptual and neural mechanisms underlying speech and music processing in real-life listening conditions by listeners coming from different language backgrounds.

Finally, building on the hypothesis that more reliable acoustic dimensions would receive more attention, I predicted that Mandarin speakers would be better at attending to pitch and experience more difficulties in ignoring pitch while attending to other acoustic dimensions compared to English speakers. In line with my expectations, I observed Mandarin speakers' advantage in attention to pitch – a benefit also achieved through musical training. I did not find that Mandarin speakers had trouble ignoring pitch, which suggests that there is no globally increased pitch salience among Mandarin speakers.

My results clearly show that the effects of L1 on the accuracy of attention are different for musicians compared to non-musicians and depend on whether pitch or duration is being measured. Regardless of these between-condition differences, a general pattern towards a positive correlation across domains and conditions has emerged. Better attention might support a more flexible use of acoustic dimensions across categorization tasks (i.e., selecting the dimension that is most reliable or suitable for a given task) – a behavioural pattern that might result from the lifetime of experience native speakers have with their L1 that allows them to learn what the most optimal dimension is. On the other hand, worse attention might lead to the inability to focus on the primary or most informative dimension and using all available information across dimensions instead. Such a strategy might emerge as a consequence of how we learn an L2 through emphasizing differences between speech tokens but without providing explicit qualitative feedback about the nature of that difference. As a result, L2 learners might rely on more global differences between the sounds than the acoustic nuances L1 speakers can perceive, or it might take them a long time and plenty of L2 input to learn how to detect them reliably.

The ability to selectively attend to specific acoustic dimensions may be advantageous in adopting successful or more native-like listening strategies. However, one significant limitation of this paradigm is that the stimuli presented during the experiment do not resemble real-life listening conditions. Most of the time, speech perception takes place in noisy environments, so cue weighting might need to be redistributed differently to account for the availability of the acoustic information in such an environment (e.g., Gordon et al., 1993; Symons, Holt, & Tierney, 2023). Further research is needed to determine the role of attention in shaping perceptual strategies in more natural conditions.

Conclusions

Overall, these results are consistent with attentional theories of cue weighting, which claim that listeners redirect their attention toward the most informative cues and those that are reliable for a given task (e.g., Francis & Nusbaum, 2002), i.e., more salient or reliable dimensions receive greater weight. However, I did not observe a substantial increase in salience of those dimensions, as evidenced by the lack of enhancements of cortical pitch tracking, leaving the question of the mechanisms underlying perceptual strategies unanswered. Furthermore, the observed behavioural patterns are not the same for native and non-native speakers, suggesting that L2 strategies in adult learners might be building up on top of the existing L1 strategies and might not have been optimized for use in the context of a new language. Whether such optimization is even possible is a topic of ongoing research. Finally, my findings further validate that certain elements of auditory performance are contingent upon both L1 background and musical experience, highlighting the distinct impact of these two types of experience on auditory perception.

Chapter 3. Individual differences in cue weighting strategies for L2 speech prosody

Abstract. Perception of speech is a complex process that depends on effective integration of acoustic information. The relative informativeness and weights assigned to any given cue are said to arise from distributional characteristics of listeners' native language where the most reliable or informative cue receives stronger weights. However, even among people from the same language background we find evidence of individual variability in employed perceptual strategies. Research to date has focused mostly on differences in strategies driven by distinct language backgrounds (e.g., showing stronger reliance on pitch by tonal language speakers compared to non-tonal speakers), not on within-group variation, and so, the sources of this variability remain largely unexplored. In this study, I aim to explore individual variability in cue weighting strategies employed by second language learners and the potential sources of that variability. I examine whether these strategies vary across tasks and attempt to reconcile whether and, if so, to what extent individual differences in language learning experience, auditory abilities, attention, and dimensional salience can explain the differences in cue weighting strategies. To achieve that goal, I measured sensitivity thresholds to pitch and duration differences, accuracy of dimension-selective attention to those dimensions and cue weighting strategies during speech and musical beats perception of 60 native Mandarin speakers. Additionally, participants' brain activity was measured with EEG during a frequency tagging paradigm to measure dimensional salience of pitch and duration. I find that early immersion in an L2 environment does not predict cue weighting strategies. Results also show that neither dimension-selective attention nor dimensional salience emerged as significant predictors of cue weighting strategies across language and musical beats categorization tasks. These results do not support attention-based theories of cue weighting in second language speech perception. However, we should consider this interpretation with caution, given that the analyses conducted in the current study yielded null results. I found shared strategies applied to word stress and musical beats categorization, but not across other tasks. That suggests that there is no uniform strategy that listeners apply across listening tasks.

3.1 Introduction

Understanding speech requires parsing the incoming signals into comprehensible units and categorizing these units into appropriate linguistic categories. However, multiple acoustic cues can describe a single contrast (e.g., pause, final lengthening and pitch reset signal intonational phrase boundaries, Yang, Shen, Li, & Yang, 2014; pitch movement, increased amplitude and duration and changed vowel quality define stressed syllables, Fear, Cutler, & Butterfield, 1995; Plag, Cunter & Schramm, 2011) and so, not all the cues will be weighted equally during speech perception. Existing models of cue weighting posit that more reliable dimensions receive more weight (Toscano & McMurray, 2010) and that reliability is assumed to be learned through a lifetime of first language (L1) exposure by guiding listeners' attention towards L1-relevant dimensions or enhancing their salience (Holt & Lotto, 2006; Francis et al., 2008). As a consequence, people optimize their listening strategies to L1 inputs leading to noticeable differences in strategies employed by speakers across languages. A substantial number of studies investigating cue weighting patterns focused on emphasizing the differences between groups from divergent L1 backgrounds (e.g., English vs Spanish, Llanos, Dmitrieva, Shultz, & Francis, 2013; English vs Korean, Kim & Cho, 2013, Kim, Clayards, & Goad, 2017; English vs Mandarin vs Russian, Chrabaszcz, Winn, Lin, & Idsardi, 2014). L1 experience has been implicated as one of the main factors shaping listening strategies – languages not only differ in which cues are the most relevant but also in the relative importance of these cues.

However, even among people from the same language backgrounds, there is large individual variability in how they weigh acoustic dimensions across tasks. For example, English speakers differ in their reliance on voice onset time (VOT) and fundamental frequency (F0) onset while distinguishing between voiced and unvoiced consonants in their native language (/da/ vs /ta/, Kong & Edwards, 2011, 2016; /ba/ vs /pa/, Shultz, Francis, & Llanos, 2012) and Japanese speakers vary in their weighting assigned to absolute and relative duration while differentiating Japanese singleton and geminate consonants (/seta/ vs /setta/, Idemaru, Holt, & Seltman, 2012). Furthermore, while comparing the reliance on two contrasting cues among second language (L2) learners, individuals systematically rely either on one dimension or another or a combination thereof (e.g., VOT vs F0 at vowel onset in detecting L2 English stop contrasts by Korean learners, Schertz, Cho, Lotto, & Warner, 2015). This behavior persists even if the acoustic cues are equally informative (Idemaru et al., 2012), which challenges the idea that the reliability of acoustic dimensions is absolute or somewhat uniform across speakers of the same L1. Instead, cue use seems to be dependent not only on the characteristics of the specific L1 but also on individual differences. Studies have also demonstrated that these differences in cue use are stable over time (e.g., individual patterns of responses are maintained over multiple testing sessions with time intervals as long as two months; Idemaru et al., 2012) and consistent within individuals (Schertz et al., 2015, 2016), indicating that these differences are not fleeting properties, but rather reliable strategies. It remains to be seen however whether individual differences in cue weighting strategies are stable across tasks. Only limited evidence exists showing that cue weights correlate across some phonetic contrasts (/bat/ vs /bet/, /Luce/ vs /lose/ and /sock/ vs /shock/), but not others (/bog/ vs /dog/ and /dear/ vs /tear/, Clayards, 2018). But so far, limited attempts have been made to explain what the source of these differences might be. It could be that while years of L1 exposure shape our listening strategies, other experiences, sensory constraints and individual differences in sound perception modulate them, which results in weightings that are not identical across listeners. It is also possible that perceptual strategies are not rigid but rather task-dependent and adaptable to changing listening conditions (e.g., adaptability to atypical accents, Jasmin, Tierney, Obasih, & Holt, 2023).

One potential explanation for these differences, at least among L2 learners, is the degree of experience a given person has with L2 learning. Research identified a range of factors that play a substantial role in defining language learning trajectories, such as the age of arrival in an L2 environment (e.g., later AOA linked to stronger L1 accent, Yeni-Komshian, Flege, & Liu, 1997; earlier AOA linked to better L2 pronunciation, Yeni-Komshian, Flege, & Liu, 2000; earlier AOA associated with greater grey matter density and white matter integrity in language-related areas, Li, Legault, & Litcofsky, 2014), length of residence in L2-speaking countries (LOR; significant improvements in L2 production during the first year of immersion, Saito, 2013; continued improvements up to 6 years of LOR, Saito, 2015a) or good quantity and quality of L2 interactions (i.e., dominance of L2 use linked to better perception and production; Flege & Liu, 2001), that might contribute to changing perceptual strategies over the course of language learning experience. Since perceptual strategies are said to develop as a function of experience with the language input regularities (Holt & Lotto, 2006), one could expect gradual shifts towards more native-like cue weighting patterns with increasing exposure to reliability and informativeness of L2 cues. However, research to date failed to deliver evidence supporting this hypothesis. For example, a study showed no significant correlations of LOR with cue weighting patterns (Idemaru et al., 2012), and there was no effect of a 1-year long immersion experience for Swedish and Finnish learners of English on their perception of /s/ vs /z/ contrast (Flege & Hillenbrand, 1986), or even long-term L2 experience on the use of duration in categorization of a non-native vowel contrast (naïve vs experienced Catalan learners of English, Cebrian, 2006). Language experience seems to be a necessary but not sufficient factor to account for the observed variability in perceptual strategies.

It is also plausible that sensory processing poses constraints on the extent to which listeners can benefit from their language experience, as various levels of perceptual acuity might serve as a bottleneck for L2 speech perception. However, the role of these abilities in shaping perceptual strategies is relatively unexplored. It appears that differences in speech perception are not linked to a loss of auditory sensitivity to the acoustic cues underlying those perceptual decisions. Miyawaki and colleagues (1975) showed that even though Japanese learners of English struggle with distinguishing between the English /r/ and /l/ contrast, they showed comparable performance to native speakers' discrimination of isolated three-formant tones. On the other hand, a fruitful line of research showed that Japanese native speakers lack sensitivity to the third formant (F3), which is generally not useful for distinguishing phonemes in their native language but is a crucial cue for resolving the contrast between the English /r/ and /l/ (Iverson et al., 2003; Ingvalson, Holt, & McClelland, 2012). Japanese learners eventually improve their speech perception and production skills, but they might not improve their perception of F3 and rely entirely on secondary cues such as F2 (Ingvalson, McClelland, & Holt, 2011) or integrate the information from F2 and F3 (Yamada & Tohkura, 1992). Similarly, native speakers of tonal languages demonstrated superior pitch perception abilities (e.g., more precise pitch discrimination, Giuliano et al., 2011; better pitch interval discrimination, Zheng & Samuel, 2018; more robust neural pitch encoding, Chandrasekaran, Krishnan, & Gandour, 2009; Bidelman, Gandour, & Krishnan, 2011) – a strategy that might not be optimal for learning languages in which pitch plays a more secondary role. One consequence of these differences in cue use might be the larger number of categorization errors. But most importantly, an increased sensitivity to irrelevant acoustic dimensions might place a strain on processing resources – detecting critical acoustic differences might require more focused attention and longer processing times.

Another possibility is that, as suggested by theories of speech perception (Francis & Nusbaum, 2002; Holt et al., 2018), prolonged exposure to L1 regularities changes the relative salience of acoustic dimensions (i.e., how noticeable or important one dimension is compared to another), resulting in shifts of attention towards the most reliable cues. According to these theories, enhanced attention to acoustic dimensions or their salience would lead to observable differences in cue weighting between speakers of different languages (i.e., more salient or attended dimensions would receive stronger weight). However, despite their prominent role in cue weighting and speech perception theories, the role of relative salience and dimension-selective attention in these processes is largely unknown. Ample evidence exists that acoustic information particularly important in listeners' L1 receives more weight. For example, pitch contour plays a vital role in conveying meaning in Mandarin Chinese and native speakers of that language were shown to rely more on pitch movements signalling lexical contrasts than Dutch speakers (Braun & Johnson, 2011), overweight pitch in phrase boundary and lexical stress categorization (Jasmin, Sun, & Tierney, 2021), and even overuse pitch when speaking (Nguyen, Ingram, & Pensalfini; 2008; Zhang, Nissen, & Francis, 2008). However, despite their theoretical grounding in attention-to-dimension theories, limited attempts were made to measure attention or salience and link them to cue weighting strategies employed by participants. To my knowledge, only two studies examined attention and showed that perceptual strategies change when people are distracted by the increased attentional load (completing arithmetical tasks, Gordon et al., 1993; Kong & Lee, 2018) or presence of competing talkers (Symons, Holt, & Tierney, 2023), suggesting that listeners adaptively shift attention in response to demanding listening conditions. Some other studies talk about participants being "more attentive" (Braun & Johnson, 2011; pp. 585) or researchers measuring "perceptual attention to acoustic cues" (Escudero, 2001; pp. 250) while talking about cue weighting, but without direct assessment of attention itself. This could potentially lead to misleading conclusions about the role of attention or salience in shaping perceptual strategies. It is a tenable hypothesis but has yet to be fully empirically tested.

Present study

The main goal of this study is to test the hypothesis about the role of relative salience and dimension-selective attention in explaining individual differences in cue weighting.

Most importantly, I attempted to reconcile a long-lasting debate about the role of relative salience and dimension-selective attention in driving cue weighting strategies. The dominating in the field hypothesis stands that L1 experience inevitably leads to shifts of attention towards the most informative and reliable cues or enhances their salience (Francis & Nusbaum, 2002; Holt et al., 2018). According to these theories, such dimensional enhancements would be reflected by assigning them stronger weights during speech perception. In the current study, I tested this hypothesis by measuring the relative attention to pitch versus duration during a behavioural dimension-selective attention task where participants needed to detect repetition in one acoustic dimension while ignoring changes in another dimension. I also measured the relative salience of pitch versus duration with the frequency tagging paradigm, a method that tracks the brain's response to changes in acoustic dimensions tagged to specific frequencies (Symons, Dick, & Tierney, 2021). I test whether these factors can reliably predict cue weighting strategies across tasks. If the hypothesis about the role of attention and salience in perceptual strategies is true, I expect these measures to emerge as significant predictors of cue weighting in the regression models.

Secondly, I attempted to link a range of experiential factors such as the age of arrival to an L2-speaking country for the first time, length of immersion in an L2 environment and daily L2 English use at participants' home and in social and professional settings to L2 learners' cue weighting strategies. Previous research offers evidence that leads to mixed conclusions. Multiple studies show that earlier AOA, longer LOR and more frequent L2 use are linked to better or more native-like performance (e.g., Yeni-Komshian, Flege, & Liu, 2000; Flege & Liu, 2001; Saito, 2013), while no correlation between immersion experience and cue weighting patterns was found (e.g., Idemaru et al., 2012). In this study, I provided a more detailed account of what constitutes a language learning experience. I not only measure the quantity of immersion in an L2-speaking environment (i.e., when it started and how long it lasted) but also its quality by surveying the daily L2 use. Building on the vast literature in L2 learning suggesting that good quality of L2 immersion experience is a necessary precondition to developing native-like strategies, I predicted these factors would emerge as significant predictors of cue weighting.

Finally, it is unknown whether individual differences in cue weighting strategies are stable across tasks. Evidence from L1 suggests that individuals with musical training can optimize their strategies for a given task due to the domain-general enhancement in their ability to selectively attend to speech characteristics (Symons & Tierney, 2021). However, it is unknown whether similar patterns can be found for L2 learners who show superior pitch processing due to their tonal language background (e.g., Pfordresher & Brown, 2009; Giuliano et al., 2011). Furthermore, if cue weighting patterns reflect differences in auditory processing abilities, this would be manifested by consistent strategies across tasks. Here, I examine the correlations between tasks to provide insight into this issue. Overall, this study aims to provide a better understanding of the relationship between attention, salience, and perceptual strategies.

3.2 Methods

Participants

Participants were the same 60 native Mandarin speakers reported in Chapter 2. They were all speaking English as a second language but were not raised bilingually. They learned English as a foreign language from primary school ($M_{\text{EFL age}}$=7.58, $SD_{\text{EFL age}}$=3.05), but they all reported only elementary instruction and limited opportunities for training at this age, and their use of English was limited to a classroom setting. Participants reported only 1 to 17 months (M_{LOR}=7.41, SD_{LOR}=3.21) of L2 English language immersion experience. They all arrived in English-speaking countries in their early adulthood (M_{AOA}=21.08, SD_{AOA}=3.52). Only one participant reported their age of arrival as 11 years old when they visited the United States with their parents. However, they were not immersed in the English-speaking community at that time (i.e., their first immersion experience was when they began their studies in the UK). Only 7 participants reported speaking an additional language (1 x Russian, 1 x French, 1 x German, 2 x Japanese and 2 x Korean). According to the criteria outlined by Zhang et al. (2020), 29 participants were considered musicians (i.e., they reported obtaining more than 6 years of systematic musical training). A summary of participants' language and musical experience is shown in Table 7.

Table 7. Summary of language background and musical training information.

	M	**SD**	**Range**
Immersion experience:			
AOA (age of arrival)	21.08	3.52	11 – 29
LOR (length of residence in months)	7.41	3.21	1 – 17
EFL learning in classroom setting:			
Age when started	7.58	3.05	3 – 18
Length of training (in years)	12.65	4.12	1 – 22
Current L2 use (%):			
In professional settings	69.08	23.35	20 – 76
In social settings	30.98	23.26	4 – 95
At home	9.03	18.47	0 – 80
Musical training (N=29)[10]:			
Age when started	7.31	3.11	3 – 18
Length of training (in years)	10.86	4.02	6 – 19

Basic demographic information about the participants' language and musical training was collected during the prescreening interview to ensure their eligibility for the study before testing. Since I was interested in examining non-native strategies, I ensured that only participants with little to no immersion experience were invited to participate[11]. To gather more detailed information about participants' language background I used a tailor-made questionnaire that captured variables relevant to successful second language learning in naturalistic (Language Contact Profile: Freed, Dewey, Segalowitz, & Halter, 2004) and classroom settings (Foreign Language Experience Questionnaire: Saito & Hanzawa, 2016).

[10] Descriptive statistics were computed only for participants who reported having more than 6 years of musical training (Zhang et al., 2020).

[11] The main goal of the project was to offer targeted training to improve participants' listening strategies (the training will be described in Chapter 5), so it was crucial to recruit only people with limited immersion experience. They all had learned English as a second language at school and had to pass a recognized English language test (e.g., IELTS, TOEFL, Cambridge English, Person English Test) to enter education in the UK. However, they had little to no experience in daily communication with native speakers of English.

Behavioural measures

Dimension-selective attention

Stimuli

The base stimuli were eight unique tokens, four verbal and four non-verbal, varying along fundamental frequency (F0, correlate of voice pitch) and duration. Speech stimuli were derived from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick & Tierney, 2020). Specifically, two versions of the phrase "Tom likes barbecue chicken", one with emphasis on the word "barbecue" and one without, were used to capture natural pitch and duration variation. The first vowel /a/ was extracted from both versions and morphed along F0 using STRAIGHT voice morphing software (Kawahara & Irino, 2005) resulting in a continuum of 100 samples varying in pitch. Other acoustic parameters were kept constant during that manipulation. Samples from Level 1 and Level 56 that differed by approximately 2 semitones (i.e., 110.88 and 124.40 Hz respectively) were chosen as base stimuli and submitted to duration manipulations. These values were chosen to guarantee balanced pitch vs duration salience across stimuli. Using Praat software (Boersma & Weenink, 2001), the duration of the vowel was morphed to approximately 70 and 175 ms. Non-verbal stimuli consisting of complex tones with 4 harmonics were generated with custom MATLAB scripts to match the acoustic properties of verbal sounds. Similarly to verbal stimuli, tones varied along two dimensions, pitch and duration, resulting in four unique tokens (110.91–124.72 Hz and 70.58–175.83 ms).

The base stimuli were concatenated into 2 Hz sequences, with pitch and duration varying at different rates (every 3 sounds = 0.67 Hz or every 2 sounds = 1 Hz). Repetitions (i.e., instances where dimension did not change at the expected rate) were inserted into half of the sequences for each dimension, resulting in four trial types: pitch repetition only, duration repetition only, repetitions in both dimensions, and no repetitions in either dimension. The stimuli for each trial type and attention condition were identical, with the focus of attention being the only varying factor. From each stimulus set, 64 stimuli were randomly selected and assigned to attend pitch or attend formant conditions, with 32 trials per attention condition. Two versions of the task (versions A and B, randomized across participants) were used, with the stimuli assigned to opposite attention conditions in each version. For instance, if a recording was assigned to attend pitch in version A, it would be assigned to attend duration in version B, ensuring that the stimuli were counterbalanced across attention conditions in both versions of the task.

Task

During the experiment, participants were presented with sequences of verbal and non-verbal sounds that varied in pitch and duration at two different rates. At the beginning of each block, they were instructed to attend to changes in one of these dimensions. Each trial started with 500 ms of silence, followed by the presentation of the stimulus. After the stimulus played, a prompt appeared on the screen, asking participants whether they heard a repetition within the attended dimension. Participants indicated their responses by clicking the 'Yes' or 'No' button on the screen. Feedback on the accuracy of the response was provided after each trial. Participants received instructions for the next block and were allowed to take a break before starting the next block.

Prior to the main task, participants were introduced to the different levels of pitch and duration and were presented with single-dimension sequences that varied at different rates with and without repetitions to familiarize them with stimuli. They then completed a short training task almost identical to the main task, with the only difference being that only the attended dimension varied while the unattended dimension was held constant. Participants needed to correctly answer at least 6 out of 8 trials (75%) to progress to the main task (up to 3 attempts were allowed). Participants who did not pass the performance threshold did not continue to the next stage of the study.

The task comprised 4 blocks for each domain (verbal and non-verbal), corresponding to 2 attention conditions and 2 rates of change. At the start of each block, participants were instructed which dimension to attend to and the expected rate of change. The final score was computed as the hit rate collapsed across the dimension change rate for each attended dimension. I computed normalized accuracy for verbal and non-verbal stimuli by dividing pitch accuracy by the sum of pitch and duration accuracy and used these normalized values for further analysis.

Discrimination tasks

The stimuli consisted of verbal and non-verbal 100-step continuums changing either in pitch or duration. The endpoints of the pitch continuum for verbal stimuli were /a/ vowels with the F0 equal to 110.88 and 124.40 Hz selected as base stimuli for the dimension-selective attention task (levels 1 and 56). STRAIGHT voice morphing software was used to create a 100-step pitch continuum between these endpoints while keeping the duration constant at 105 ms. For the duration discrimination stimuli, Praat software was used to create a 100-step duration continuum by modifying the unstressed version of the vowel from 70 ms to 175 ms while keeping the pitch constant. The non-verbal stimuli were complex tones with four harmonics and 10-ms linear on and off ramps. The pitch and duration continuums were created with custom MATLAB scripts to match the properties of the speech stimuli. The pitch discrimination continuum consisted of 100 stimuli with a range of F0 from 110.88 Hz to 124.40 Hz, while the duration discrimination continuum consisted of 100 stimuli with a range of duration from 75 ms to 175 ms, with a constant F0 of 110.88 Hz.

During each trial, participants were presented with three stimuli in AXB format, where X matched either stimulus A or B. Participants were required to identify whether the first or third sound was different from the other two by pressing the corresponding number (1 or 3) on the screen. The stimuli were presented with a constant interstimulus interval of 500 ms. An adaptive three-alternative forced-choice procedure was used, based on the transformed up-down procedure described by Levitt (1971). The task difficulty increased after every second correct response and decreased after every incorrect response. The presentation started at stimulus level 50 with an initial step size of 10. After the first reversal, the step size decreased to five, and after the second and third reversals, it decreased to two and one, respectively. This step size remained constant until the end of the presentation, allowing the detection of the smallest difference between stimuli that participants could perceive. There were four blocks in total representing the stimulus domains (verbal and non-verbal stimuli) and acoustic dimensions (pitch and duration). The order of the tasks was randomized and balanced across participants using a Latin square design. Each block was terminated after 70 trials or seven reversals, and the final score was calculated as the levels of each reversal from the second onward. A lower threshold level indicated better performance. To capture the variations in sensitivity thresholds for pitch vs duration, I computed the relative inverse variance in representation of the two dimensions (Ernst & Banks, 2002). The log-transformed thresholds were used to compute the pitch-duration variance estimates for verbal and non-verbal stimuli using the formula:

$$\hat{S} = \frac{1/(log(pitch\ threshold)^2)}{1/(log(duration\ threshold)^2)}$$

Values closer to 1 represented better pitch discrimination ability, whereas values closer to 0 represented better duration discrimination ability.

Cue weighting categorization tasks

Stimuli

The stimuli comprised of sentences exemplifying three distinct prosody features (i.e., linguistic focus, phrase boundary, and lexical stress) derived from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick & Tierney, 2020) and musical beat patterns (Jasmin, Sun, & Tierney, 2021). All speech samples were recorded by a native Southern British English voice actor. The stimuli were created from recordings of two sentences or words with contrastive prosody. The linguistic focus stimuli were created from the sentences "Dave likes to STUDY music, but he doesn't like to PLAY music" and "Dave likes to study MUSIC, but he doesn't like to study HISTORY". The phrase boundary stimuli were the sentences "If Barbara gives up, the ship will be plundered" and "If Barbara gives up the ship, it'll be plundered", and the lexical stress stimuli were the word "compound", pronounced with stress on either the first or second syllable. To create the stimuli, identical portions of each recording were extracted, resulting in two versions of the same phrase with the prosodic contrast located at different positions. These two versions of each sample were then morphed using the STRAIGHT software (Kawahara & Irino, 2005; Jasmin et al., 2020) that involved the orthogonal manipulation of F0 and durational morphing rates to create the stimuli continua. The musical beat stimuli comprised of 18-tone sequences, where each beat featured six tones repeated three times. The tones were four-harmonic complex tones with uniform amplitude across harmonics, and 15-ms on/off cosine ramps. The pitch and duration of these tones varied across four levels and indicated either a two-note grouping (“strong—weak” pattern, march time) or a three-note grouping (“strong—weak—weak” pattern, waltz time). The strength of the groupings was determined by the increased pitch or duration of the first tone in the two- or three-note grouping relative to other tones. Both verbal and non-verbal stimuli sampled a 4-by-4 acoustic space across duration and F0 (see Figure 3 in Chapter 2 presenting the stimulus grids).

Procedure

During each of the four categorization tasks, participants were presented with stimuli that varied orthogonally in the degree to which fundamental frequency (F0) and duration indicated one of two possible linguistic interpretations. Following the presentation of each stimulus, they were asked to categorize it as belonging to one of two categories: phrase with early or late closure ("If Barbara gives up, the ship" vs "If Barbara gives up the ship"), emphasis on the first or second word ("STUDY music" vs "study MUSIC"), lexical stress on the first or second syllable ("COM-pound" vs "com-POUND"), and musical beats occurring either every two or three notes ("strong—weak" vs "strong—weak—weak" patterns). Before the main task, participants were given examples of each linguistic contrast with unaltered pitch and duration and two practice trials with written feedback. The main tasks were identical to the practice trials, except that feedback was no longer provided, and all 16 stimuli were presented in random order. There were 10 blocks of each categorization task, which were interleaved in the following order: musical beats, linguistic focus, lexical stress, and phrase boundary. Practice trials were included only in the first block of each task. Participants were provided with progress updates after completing each block.

Analysis

Since the distribution along one cue likely affects the weight of the other cue (Schertz & Clare, 2019), I computed normalized cue weights as a relative measure of pitch versus duration for all categorization tasks. I used Firth's biased-reduced logistic regression method (Firth, 1993) to estimate pitch and duration levels (from 1 to 4) for each participant that predicted the binary responses during each categorization task. The coefficients obtained for pitch contour and duration were then normalized such that their sum was equal to one (for a similar approach see Holt & Lotto, 2006; Idemaru et al., 2012; Jasmin et al., 2020). The resulting normalized perceptual weight ranged between 0 and 1, where values closer to 1 indicated a higher reliance on pitch contour than duration, whereas values closer to 0 indicated the reverse, and 0.5 indicated equal reliance on both features. I performed the analysis using Firth's regression implemented in the logistif R package (Heinze et al., 2022).

Neural measures

EEG data acquisition

EEG data was recorded from 32 Ag-Cl active electrodes using a Biosemi™ ActiveTwo system with the 10/20 electrode montage and two reference electrodes placed on earlobes. Data were recorded at a 16,384 Hz sampling rate and digitized with a 24-bit resolution. Impedance was kept below 20 kΩ throughout the testing session. Triggers marking the beginning of each trial (every 6 tones, or 1.2 seconds) were recorded from trigger pulses and sent to the data collection computer. All EEG data processing and analysis steps were carried out in MATLAB (MathWorks, Inc) using the FieldTrip M/EEG analysis toolbox (Oostenveld et al., 2011) in combination with in-house scripts.

Frequency tagging paradigm

Stimuli

The base stimuli were identical to those used in the dimension-selective attention task and included speech and tones sampling a 2 (pitch) x 2 (duration) space. I created 5-Hz sequences (96 seconds long) where pitch and duration varied at fixed rates (1.67 Hz or 2.5 Hz). 20 repetitions were inserted into each sequence to prevent the stimuli from becoming overly predictable. The amplitude of 3-5 randomly selected stimuli (32 in total) was reduced by 25% (-12.04 dB) to create amplitude oddballs. Oddball timing was randomized in each sequence, with the exception that oddballs could not occur in the first or last 4.8 seconds (4 epochs) of the sequence and could not occur within 4.8 seconds of another oddball. The same sequences were presented to all participants, with the order of presentation counterbalanced across participants. Stimuli were presented diotically via insert earphones ER-3A (Etymotic Research, Elk Grove Village, IL) at a maximum of 80 dB SPL and a sampling rate of 44100 Hz using PsychoPy3 (v 3.2.3).

Behavioural task

Participants listened to the sound sequences and responded with keyboard presses to occasional quiet tones. The purpose of this behavioral task was to maintain participants' attention, without directing their attention to any specific acoustic dimension. Prior to the main task, participants completed a brief practice run to familiarize themselves with the task. They listened to sequences of verbal and non-verbal sounds for about a minute each and continued until they reached at least 5 out of 6 correct responses, without making too many errors. During the practice, feedback was displayed on the screen, indicating the number of correct and incorrect responses, and missed targets. Most participants completed the practice successfully on their first attempt, while the remaining participants repeated the practice block to ensure good performance during the EEG recording. The main task was identical to the practice, but with longer sound sequences and no visual feedback. Behavioral performance was monitored to ensure that participants remained focused. The task comprised four blocks, each containing four two-minute sound sequences.

Intertrial phase coherence (ITPC)

The recorded data were down-sampled to 512 Hz and re-referenced to the average of the earlobe reference electrodes. A low-pass zero-phase sixth-order Butterworth filter with a 30 Hz cutoff was then applied, followed by a high-pass fourth-order zero-phase Butterworth filter with a cutoff of 0.5 Hz. The data were then segmented into epochs of 1.2 seconds based on the recorded trigger pulses. Independent component analysis (ICA) was performed to correct for eye blinks and horizontal eye movements. Components corresponding to these artifacts were visually identified and removed. Subsequently, data were extracted from the 9 channels with the maximum ITPC (AF4, F3, Fz, F4, FC1, FC2, FC5, Cz, C3) across all participants (N=60), resulting in a cluster of frontocentral channels. The data were averaged across the selected channels. Any remaining artifacts exceeding +/-100 µV were rejected. To assess cortical tracking of acoustic dimensions, inter-trial phase coherence (ITPC) was calculated at the frequencies of dimension change. This was done by applying a Hanning-windowed fast Fourier transform to each 1.2-second epoch, converting the complex vector at each frequency to a unit vector, and averaging across trials. The resulting length of the average vector provided a measure of phase consistency, which ranged from 0 (no phase consistency) to 1 (perfect phase consistency). The degree of ITPC at the frequency tagged to a certain dimension provided an index of dimensional salience. I computed normalized salience for verbal and non-verbal stimuli by dividing pitch tracking by the sum of pitch and duration tracking and used these normalized values for further analysis.

General procedure

Prospective participants who responded to recruitment adverts underwent a preliminary screening via an individually scheduled brief telephone or video call with the researcher. The purpose of the call was to ask questions about participants' demographics, language, and musical background to ensure they met all the inclusion criteria. After a short, guided interview, the researcher provided them with detailed explanation of the experimental procedure and task instructions and addressed participants' questions about the study and its purpose. Eligible participants were given consent forms, and upon granting consent they received links to three sets of online tasks (Parts 1, 2 and 3). All tasks were designed and hosted on the Gorilla Experiment Builder platform (Anwyl-Irvine et al., 2020). Part 1 included a detailed demographics questionnaire, discrimination tasks and dimension-selective attention tasks. Part 2 consisted of a series of categorization tasks and Part 3 of language assessment tasks. Upon completion of all the online tasks participants were invited to the lab to record their brain activity with EEG (i.e., Parts 1, 2 and 3 and EEG session constitute the Pre-Test measurement). After attending the EEG session, participants completed several days of online training and completed Parts 1, 2 and 3 tasks and returned to the lab for a second EEG session in the lab (i.e., Post-Test). The second session took place approximately two weeks after the first measurement[12]. I used the post-

[12] A crucial element of the testing schedule was to ensure that participants complete the second measurement immediately after the training. However, I allowed some flexibility

test data to establish the reliability of measures of interest for this Chapter and the remaining data collected during these stages will be discussed in detail in Chapters 4 and 5. The main data analyzed in this Chapter includes the behavioural and neural data from Mandarin speakers at Time 1 (i.e., at Pre-Test) as outlined in Figure 1. Processed data and analysis scripts can be found at: *https://osf.io/5j8pe/?view_only=874b45bef48c4839807867f12a02809a*.

Data collection was conducted at the Department of Psychological Sciences at Birkbeck, University of London and all the ethics procedures were approved by the departmental Ethics Committee. To compensate for their time, all participants received cash reimbursement of £10 per hour or its equivalent in course credits.

3.3 Results

Reliability of auditory sensitivity, dimension-selective attention, and dimensional salience measures

I used Pearson's correlation coefficient as a measure of the consistency of results between the first and second measurement of auditory sensitivity, dimension-selective attention, and dimensional salience (Table 8). Dimension-selective attention (r_{verbal}=.467 p<.001; $r_{\text{non-verbal}}$=.527, p=<.001) and dimensional salience (r_{verbal}=.348, p=.006; $r_{\text{non-verbal}}$=.305, p<.018 showed adequate test-retest reliability. Contrary to the recent report showing good reliability of discrimination thresholds (Saito & Tierney, 2022), measures of relative perception of pitch vs duration in verbal and non-verbal stimuli in this study were found to be unreliable. Only attention and salience measures were included in further analysis.

regarding how much time elapsed from completing the first measurement until they commenced the training.

Table 8. Summary of reliability analysis for relative measures of auditory sensitivity, dimension-selective attention, and dimensional salience.

				95% CI	
Task	**Measure**	**r**	**p**	**Lower**	**Upper**
Discrimination thresholds					
	Verbal (log)	-.085	.518	-.332	.172
	Non-verbal (log)	.127	.335	-.131	.369
Dimension-selective attention					
	Verbal (normalized)	.467	**<.001**	.242	.644
	Non-verbal (normalized)	.527	**<.001**	.315	.689
Dimensional salience					
	Verbal (normalized)	.348	**.006**	.103	.553
	Non-verbal (normalized)	.305	**.018**	.055	.518
Cue weighting					
	Linguistic focus (normalized)	.340	**.008**	.095	.547
	Lexical stress (normalized)	.492	**<.001**	.272	.663
	Phrase boundary (normalized)	.729	**<.001**	.583	.830

The role of language experience, dimension-selective attention, and dimensional salience in shaping cue weighting strategies

To investigate whether experiential factors such as the age of arrival (AOA), length of residence in L2-speaking countries (LOR), and daily use of L2 English (L2 use), and relative measures of dimension-selective attention and dimensional salience (pitch versus duration) can explain the observed differences in cue weighting patterns, I conducted a series of linear regressions using the lmer4 R package (Bates et al., 2015). The outcomes of the four separate models were normalized cue weights across categorization tasks: linguistic focus, phrase boundary, lexical stress, and musical beats. As a first step (Model 1), language experience factors (AOA, LOR and L2 use) were entered into the model to establish whether they explain the differences in cue weighting patterns. Relative attention to pitch vs duration for speech and tones, as well as relative salience of pitch and duration in speech and tones stimuli, were entered into the regression model as a second step (Model 2). The likelihood-ratio test was conducted to test whether adding additional predictors improved the model fit. As summarized in Table 9, none of the predictors were found to explain the observed differences in cue weighting strategies.

Table 9. Results from linear regressions models predicting L2 cue weighting strategies based on language experience, dimension-selective attention and dimensional salience. AOA – age of arrival (i.e., the age at which participants arrived in the L2-speaking country for the first time); LOR – length of residence in L2-speaking countries.

	Model 1				**Model 2**			
Linguistic focus	**β**	**SE**	**t**	**p**	**β**	**SE**	**t**	**p**
(Intercept)	.913	.009	101.059	<.001	.913	.009	97.817	<.001
AOA	.036	.021	1.683	.098	.036	.023	1.586	.119
LOR	-.023	.021	-1.059	.294	-.021	.022	-0.966	.339
L2 use	.016	.019	.850	.399	.013	.019	.673	.504
Attention (verbal)	--	--	--	--	-.012	.021	-.543	.590
Attention (non-verbal)	--	--	--	--	.012	.021	.552	.583
Salience (verbal)	--	--	--	--	<.001	.020	.006	.995
Salience (non-verbal)	--	--	--	--	.016	.020	.833	.409
	Model 1				**Model 2**			
Phrase boundary	**β**	**SE**	**t**	**p**	**β**	**SE**	**t**	**p**
(Intercept)	.411	.036	11.496	<.001	.411	.037	11.173	<.001
AOA	-.028	.084	-.337	.738	-.010	.090	-.108	.802
LOR	-.040	.085	-.473	.638	-.040	.088	-.439	.639
L2 use	-.044	.073	-.598	.552	-.039	.077	-.507	.467
Attention (verbal)	--	--	--	--	-.021	.085	-.252	.808
Attention (non-verbal)	--	--	--	--	.040	.084	.472	.914
Salience (verbal)	--	--	--	--	.058	.080	.733	.662
Salience (non-verbal)	--	--	--	--	-.019	.078	-.244	.614
	Model 1				**Model 2**			
Lexical stress	**β**	**SE**	**t**	**p**	**β**	**SE**	**t**	**p**
(Intercept)	.899	.008	107.591	<.001	.899	.009	105.152	<.001
AOA	.036	.020	1.840	.071	.037	.021	1.791	.079
LOR	-.024	.020	-1.197	.236	-.021	.020	-1.009	.318
L2 use	-.001	.017	-.007	.994	-.004	.018	-.197	.845
Attention (verbal)	--	--	--	--	.002	.020	.093	.926
Attention (non-verbal)	--	--	--	--	-.009	.020	-.454	.652
Salience (verbal)	--	--	--	--	.008	.018	.438	.664

Salience (non-verbal)	--	--	--	--	.016	.018	.895	.374

	Model 1				**Model 2**			
Musical beats	**β**	**SE**	**t**	**p**	**β**	**SE**	**t**	**p**
(Intercept)	.848	.021	39.791	<.001	.848	.022	39.233	<.001
AOA	.077	.050	1.538	.130	.075	.053	1.426	.160
LOR	-.006	.051	-.126	.900	-.008	.052	-.168	.867
L2 use	.071	.044	-1.638	.107	-.075	.045	-1.664	.102
Attention (verbal)	--	--	--	--	-.037	.050	-.736	.465
Attention (non-verbal)	--	--	--	--	.074	.050	1.499	.140
Salience (verbal)	--	--	--	--	-.004	.047	-.081	.936
Salience (non-verbal)	--	--	--	--	.032	.06	-.705	.484

Consistency of cue weighting strategies across tasks

I used Spearman's correlations to examine the relationship across strategies (data were not normal according to the Shapiro-Wilk test, p<.05). Participants' cue weights were moderately correlated between lexical stress vs musical beats (ρ.=.37, $p_{\text{FDR-corrected}}$=.005; Table 10 and Figure 14).

Table 10. Correlation table for normalized cue weights across categorization tasks. All corrected for multiple comparisons with the FDR correction. Significant correlation is highlighted in bold.

	Linguistic focus	Lexical stress	Phrase boundary	Musical beats
Linguistic focus	--	--	--	--

Lexical stress	ρ=.351
	p=.210
		--	--	--

Phrase boundary	ρ=-.111
	p=.517
		ρ=-.110
		p=.517
			--	--

Musical beats	ρ=.106
	p=.750
		ρ=.366
		p=.005
			ρ=-.083
			p=.517
				--

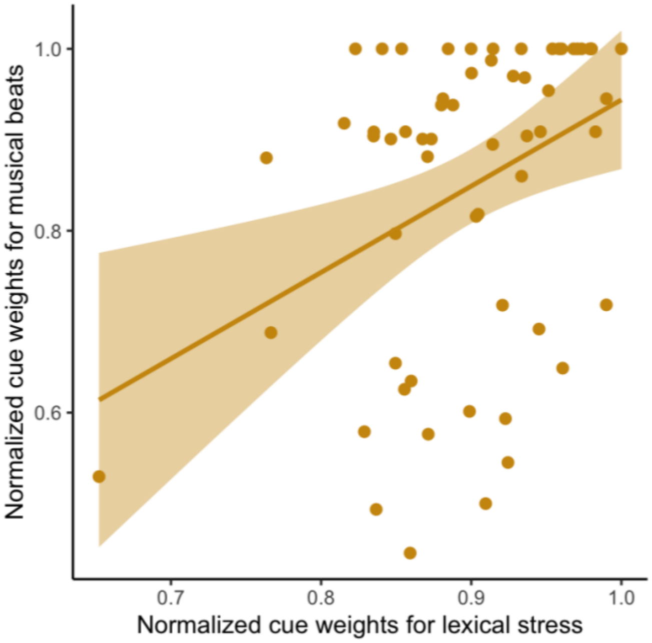


Figure 14. Scatterplot showing relationship between normalized cue weights for lexical stress and musical beats. Shaded area represents 95% confidence interval.

3.4 Discussion

In this study, I set out to examine whether the hypothesis concerning the influence of relative salience and dimension-selective attention on cue weighting holds true. I assessed behavioural indices of auditory sensitivity thresholds and attention, as well as neural tracking of acoustic dimensions as a measure of dimensional salience and used regression models to test this hypothesis. None of the factors predicted by the speech perception theories as shaping cue weighting strategies played a role in explaining the observed differences. I also examined whether individual differences in cue weighting strategies are stable across tasks. I found a significant correlation between normalized cue weights for lexical stress and musical beats but not for other tasks.

Perceptual strategies do not change during early L2 immersion experience

No evidence was found that the L2 language learning experience, as exemplified by participants' immersion age of onset and length and daily English use, has any influence on cue weighting strategies across language and musical beats tasks. A likely explanation of such a lack of effect is that compared to a lifetime of L1 experience, the L2 immersion experience was rather short, limited in its scope and did not offer as much listener-directed language input as that to which listeners are exposed when acquiring their native language (i.e., they are still L2 beginners; Saito, 2015b). Throughout development, listeners are exposed to mass amounts of speech that can serve as a solid base for tuning the auditory system to L1-relevant contrasts and extracting statistical regularities (Holt & Lotto, 2006) that cannot be easily overturned by 18 months of immersion in an L2 environment.

There is a possibility that the acquired strategies slowly change as a function of accumulated L2 exposure. However, these effects might be marginal and modulated by a range of biological and cognitive factors (Turker, Seither-Preisler, & Reiterer, 2021; Turker & Reiterer, 2021), making measuring these changes a rather tricky task. It is also not out of the question that these effects are hard to capture because of the crude methodology we use to measure them. For example, self-reports of daily L2 use can omit significant details about the quality of the reported interactions (e.g., bi-directional communication with native L2 speakers would likely be more beneficial for learners than passive listening or listening to L1-accented speech). Similarly, we do not have any insight into what L2 classroom teaching methods participants were exposed to and what effect this had on initiating any changes to perceptual strategies that might have occurred before their arrival to the UK (e.g., did they obtain any training in pronunciation, what was the level of exposure to foreign-accented vs native-like L2 English pronunciation). If L2 learning took place in an L1-accented setting, this might have cemented the application of their L1 strategies in the L2 context. Unfortunately, we cannot answer that question unless we conduct more systematic longitudinal measurements of cue use, auditory processing, and sound encoding across the lifespan.

Perceptual strategies might be task-dependent

While comparing strategies across tasks, I observed that cue weighting patterns for lexical stress and musical beat categorization were correlated among Mandarin speakers.
A common characteristic of both stimuli is their rhythmical pattern, as both can be described as sequences of strong vs weak notes or syllables. Rhythm has essential functions in both music and speech (e.g., segmentation; Cutler, 1990; for review see Cason & Schon, 2012). For example, rhythm perception abilities were shown to explain 47% of variance in morpho-syntactic competence (Gordon et al., 2014), and deficits in rhythm processing and timing were associated with a variety of language learning problems (e.g., dyslexia, Leong & Goswami, 2014; specific language impairment, Cumming, Wilson, & Goswami, 2015; developmental language disorders, Ladanyi et al., 2023). Therefore, it is possible that during categorization of stress and beats, listeners detect the underlying rhythms and rely on a strategy that works across structurally similar patterns since the musical beat stimuli adhered to the same pattern as lexical stress stimuli (emphasis placed on strong notes or syllables). Mandarin listeners could also employ similar strategies to process these two types of stimuli, as they can be distinguished based on the same primary cue. It is worth noting, however, that in English, linguistic focus is also conveyed by patterns of emphasis placed at different levels of linguistic complexity (syllable vs word level), but participants did not use the same strategy across all three tasks. These similarities might be obscure to non-native speakers, and less experienced L2 learners might not use this information to their benefit. It is also possible that Mandarin listeners do

not place enough importance on acoustic changes happening at slower time scales or are distracted by syllable-by-syllable modulations to which they have grown used to attending. Overall, these results highlight that no uniform strategy is applied to all stimuli. Rather, listeners optimize their cue use for a given task, depending on their familiarity and experience with available acoustic cues.

The lack of consistency in perceptual strategies across categorization tasks indicates that they are not driven by domain-general abilities such as auditory sensitivity. This is consistent with the literature showing great variability in auditory thresholds across people and suggesting that these abilities might be somewhat independent (Kidd, Watson, & Gygi, 2007). For example, listeners might excel at discriminating differences along one acoustic dimension but be less sensitive to changes along another dimension or even be equally good or bad at discriminating both. Such diverse sensitivity profiles likely have differential effects on sound perception more generally and could have helped shape listening strategies throughout the lifespan, although in a more indirect way. One other factor that might contribute to prosody perception is the ability to maintain auditory patterns in working memory. Individuals diagnosed with amusia present with deficits in pitch recognition (Peretz, 2016) and memory (Tillmann, et al., 2009). It is possible that due to their inability to memorize pitch patterns that are distributed across time for prosody, they down-weight the importance of pitch in speech perception (Jasmin, Dick, Holt, & Tierney, 2020). It appears that, instead of relying on a task-by-task availability of detectable cues, amusics use the information about cue reliability that emerged as a consequence of a lifetime of experience with unreliable pitch information (see also Holt & Lotto, 2006 for evidence of response bias contradictory to cues reliability). Similarly, poor ability to memorize pitch or duration patterns among L2 learners could have an effect on what dimensions listeners perceive as reliable. Perhaps this reliability is not, as previously suspected, a global L1-dependent reliability that applies to everyone, but a reliability that is unique to each individual and its abilities, groomed throughout a lifetime of personalized experience with the auditory world.

Perceptual strategies may rely on factors beyond attention and salience

Most surprisingly, I find that neither dimension-selective attention nor dimensional salience predicts cue weighting strategies. While this null result is a consequence of low reliability of the data, we must consider that some other factors might also contribute to explaining differences in cue weighting. Since people distracted by the increased attentional load (completing arithmetical tasks, Gordon et al., 1993) or the presence of competing talkers (Symons, Holt, & Tierney, 2023) might adaptively shift attention in response to demanding listening conditions, it is possible that cognitive abilities that reduce distractibility and enhance attentional focus (e.g., processing speed, Hui & Godfroid, 2021; working memory, Francis & Nusbaum, 2009; auditory-motor integration, Kachlicka, Saito, & Tierney, 2019; rhythm perception, Slater et al., 2018) could be potentially relevant in shaping perceptual strategies. For example, increasing working memory workload was shown to impede speech recognition in terms of reaction time and accuracy (Francis & Nusbaum, 2009), so it is plausible to assume that better working memory could have some beneficial effects.

A note of caution is also due here to acknowledge the rather mediocre overall reliability of used measures as a substantial limitation of this study. If the tasks did not measure the intended constructs in a reliable or accurate way, perhaps I could not detect the existent relationships between dimension-selective attention, dimensional salience, and cue weighting strategies. It is also possible that the relative measures I used here obscure individual differences that could be present. For example, choosing responses mainly at the two endpoints of the scale or corners of the stimuli space (i.e., categorical pattern of responses) and sampling more uniformly from the stimulus space (i.e., more gradient pattern of responses indicative of less clear boundaries between categories) would represent dramatically different listening strategies, yet they might result in the same normalized cue weights since the individual contributions of each dimension and choices across the stimuli space are somewhat blended together (Schertz & Clare, 2019). In a similar fashion, when averaging responses across the whole space, we might be diluting the differences present in the corners of the space where two cues conflict. If a listener has a strong bias towards pitch, their responses might be less accurate when pitch contribution to the linguistic contrast of interest is not clear enough (e.g., the amount of pitch is reduced compared to another dimension). It would be informative to look at patterns of responses where the two acoustic cues of interest conflict. The issue of gradience versus categoricality should be explored in more detail by looking at the slopes of categorical responses across tasks, an approach successfully applied to investigating categorical boundaries between various stimuli, including speech (e.g., Schertz, Carbonell, & Lotto, 2019).

We also must emphasize that combinations of cues in cue weighting cannot be interpreted independently, as they covary consistently with one another (i.e., dimensional integrality; Francis, Kaganovich, & Driscoll-Huber, 2008). In other words, changes implemented along one dimension lead to changes across other dimensions. Even if the acoustic cues of interest are controlled by morphing procedures, it could be that these changes disturb the natural use of weights during speech perception. Considering the possibility that the perceived reliability of acoustic dimensions might vary across individuals, this might mean that even in highly controlled experimental designs where researchers carefully choose the stimuli to balance the perceptual distances between categories, these distances might not be perceptually equivalent to people. Ultimately, it matters which dimensions were chosen for comparison as they not only interact with one another but also the relative differences between artificially set levels of stimuli might not be perceived equally by all listeners.

Finally, there is also a possibility that the experimental conditions deviate too much from everyday listening environments to elicit the naturalistic cue weighting behavior we intend to capture. Listeners develop their cue weighting strategies in a noisy world, so the strategies we measure in sterile lab conditions might artificially force participants to use strategies that have nothing to do with how they perceive sounds during more naturalistic listening (e.g., Gordon et al., 1993). A recent study showed that strategies indeed differ when listening to speech in quiet compared to noisy conditions (Symons, Holt, & Tierney, 2023).

A detailed research program that evaluates cue weighting strategies in more natural listening conditions is very much needed. More naturalistic listening paradigms could include scenarios in which target speech samples are embedded in background noise (e.g., coffee shop multi-talker environment), containing some signal distortions (e.g., telephone or internet connection distortions), or involve examining cue weighting within multidimensional and multimodal stimuli (e.g., listening to music, watching movies) where other sources might be obstructing some of the available acoustic information or compete for attention – the processes that ultimately shape perceptual strategies. It is perhaps impossible to evaluate cue weighting of all possible acoustic cues in naturalistic conditions while accounting for individual differences, but certainly some steps can be taken to make the differences we want to measure and their sources less elusive.

Conclusions

Despite the fact that my findings did not provide evidence in support of the hypotheses about the involvement of attention and salience in cue weighting, this study illuminates the need for further research in this domain to uncover what drives individual variability in perceptual strategies. I found shared strategies applied to word stress and musical beats categorization, but not across other tasks. That suggests that there is no uniform strategy that can be applied to any listening task, but rather there are strategies that draw on the acoustic resemblance of stimuli. Individual differences have important implications in various contexts. For example, inefficient perceptual strategies might lead to problems with extracting relevant information for learning new words and expressions. Without explicit training targeting the underlying issue, learners might continue reinforcing the use of suboptimal strategies hindering their progress. Inadequate strategies can also disrupt everyday communication, as they can obscure interlocutors' intended meaning and render interactions with other L2 speakers unsuccessful. Understanding these individual differences and their sources is therefore essential for supporting efficient L2 speech perception and L2 language learning more generally.

Chapter 4. Learning experience, attention and salience and their role in predicting L2 prosody acquisition

Abstract. The ability to communicate in a foreign language has become a valuable asset, yet despite the considerable effort learners put into mastering a new language, learning outcomes still vary significantly. Individual variability is also observed across a range of auditory skills, such as auditory processing, attention to acoustic dimensions or precision of neural sound encoding. Since speech is a primary source of language for most learners, these abilities may contribute to observed differences in proficiency. Indeed, some of these factors were implicated as potential predictors of second language (L2) learning success, whereas the contribution of others, such as dimension-selective attention or salience, remains unknown. Furthermore, most studies focus on explaining the drivers of overall L2 proficiency while omitting the role these factors might play in developing specific aspects of L2 proficiency, such as for example speech prosody. In this study, I test whether and to what extent the contributions of individual differences in L2 learning experience, auditory abilities and dimensional salience can explain variability in perception and production of L2 speech prosody. I also test whether any of these factors play a role in predicting learning outcomes across several other aspects of language competence: segmental perception and lexical and grammatical knowledge. Participants were 60 Mandarin Chinese learners of English studying in the UK. I surveyed their L2 learning experience, measured behavioural performance on a range of auditory tasks (sensitivity thresholds, dimension-selective attention accuracy, and cue weighting strategies) and recorded their brain activity with EEG during a frequency tagging paradigm to measure dimensional salience and during listening to continuous speech to assess cortical tracking of speech envelope. I also assessed various aspects of participants' L2 language performance (vocabulary, grammar, speech perception and production) as outcome measures. These findings provide support for cue weighting theories implicating the role of attention in L2 speech acquisition. I show that attention to acoustic dimensions emerged as a significant predictor of not only vowel and prosody perception, but also grammar knowledge, extending its role from spoken language to syntax.

4.1 Introduction

Effective communication is essential in our daily lives, but it can be a challenging endeavor for second language (L2) learners. Despite their best efforts, learners often struggle to understand L2 speech and speak with distinct foreign accents. Difficulties in acquiring a new language stem from the multifaceted nature of this process, which involves the acquisition and integration of various linguistic components. Learners need to acquire a high volume of new vocabulary (Wang & Treffers-Daller, 2017), master grammatical and syntactical rules (Su, 2001), and be sensitive to contextual differences in meaning to be able to communicate effectively. Yet another challenge is to master the spoken form of the new language. Difficulties in speech perception might arise partly because, during L2 learning, listeners need to accommodate their ears to new stimulus characteristics that often differ from sounds present in their first language (L1). Speaking an L2 in an intelligible, comprehensible (Crowther, Holden, & Urada, 2022), and fluent manner (De Jong, Groenhout, Schoonen, & Hulstijn, 2013) with a native-like accent (Derwing, Munro, & Wiebe, 1998) requires authentic production of individual sounds of that language (e.g., Rogers & Dalby, 2005) and its distinctive prosody (e.g., importance of speaking rate for perceived accentedness and comprehensibility, Munro & Derwing, 2002; pitch range and word stress for perception of L2 accent, Kang, 2010; pause location for perceived fluency; Kahng, 2017). However, despite acknowledging the complexity of language learning, our understanding of what factors can support the efficiency of that process and predict its success is still limited.

At the heart of any skill refinement lies the fundamental principle that practice makes perfect, emphasizing the role of consistent effort and repetition in achieving proficiency. In L2 acquisition, the length and intensity of learning experience go in tandem with better communication skills. Earlier age of arrival (AOA) in an L2 environment leads to better L2 pronunciation (Yeni-Komshian, Flege, & Liu, 2000), whereas later AOA leads to a stronger L1 accent (Yeni-Komshian, Flege, & Liu, 1997). AOA was also shown to have effects on various aspects of L2 speech prosody (Huang & Jun, 2011). Mandarin learners of English with later AOA had slower speech and articulation rate compared to native speakers and learners with earlier AOA. They inserted pitch accents at inappropriate locations and used phrasal breaks excessively, suggesting that adult learners have trouble processing prosodic structure of speech. Length of immersion experience in L2 environment was also shown to significantly boost L2 production proficiency in the first year of immersion (Saito, 2013) and these benefits continued in the following years (up to 6 years; Saito, 2015). Correlation with pronunciation improvements was observed for LOR as short as 3-10 months (Danish learners of English; Hojen, 2019). However, self-reported proportion of L2 English use was related more strongly to pronunciation improvements than the LOR. Other studies have also demonstrated that dominance of daily L2 use was linked to better perception and production (Chinese learners of English; Flege & Liu, 2001) and that people who continue using their L1 more than L2 speak with stronger foreign accents (Italian learners of English; Piske, MacKay, & Flege, 2001). Good quantity and quality of L2 interactions is therefore necessary for developing L2 proficiency, but alone can hardly explain the observed variance in L2 learning outcomes.

Since speech is a primary source of language for most learners and is especially important in the immersion context, it is possible that learning a new language poses stronger demands on auditory perception. Previous research indicated that during the first few years of immersion in the L2 environment, auditory processing was a stronger predictor of L2 success than AOA or LOR (Kachlicka et al., 2019, Saito, Sun, & Tierney, 2019). In the context of Polish learners of English, speech perception was best predicted by spectral psychoacoustic measures and additional contributions from auditory synchronization and neural encoding precision of the first and second formant (F1 and F2) as measured with the frequency-following response (FFR; Kachlicka et al., 2019). Another study corroborated these results, showing that accurate production of consonants and vowels by Chinese learners of English immersed in the UK for 1-9 months was predicted by phonemic coding and phase-locking at F1 and production of word stress by phase locking at F1, whereas intonation was predicted by recent L2 use inside the classroom (Saito, Sun, & Tierney, 2019). Encoding precision and auditory reproduction also explained individual differences in participants' pronunciation (adult Polish learners with LOR = 0.1-19 years; Saito, Kachlicka, Sun, & Tierney, 2020). These findings suggest that adult learners might struggle to learn an L2 because they lack the auditory precision needed for detecting subtle phonetic and prosodic differences of a new language. Further studies demonstrated that domain-general auditory processing (discrimination, reproduction and fidelity of neural sound encoding) is indeed a critical determinant of L2 learning success in adulthood across various language groups (Polish, Chinese, and Spanish native speakers) with varied lengths of residence (Saito et al., 2022). Taken together, these results not only emphasize the importance of auditory processing in L2 learning but also show that various aspects of auditory processing and other experiential or cognitive factors might play a role in explaining variance in L2 performance across a range of language measures.

Robust auditory processing skills are particularly important for L2 learning, which poses additional difficulties compared to learning an L1 due to differences between languages in how linguistic structure is conveyed by acoustic characteristics. For example, in tone languages, changes in pitch alter the lexical meaning of words, while in English, pitch plays a more secondary role, helping define prosodic structures such as phrase boundaries or word emphasis alongside other acoustic cues such as duration and intensity. Acoustic dimensions that are important for conveying structure in an individual's L1 are said to become particularly salient or likely to capture attention (Francis & Nusbaum, 2002; Holt et al., 2018). This acquired salience of the language-specific dimensions might lead to upweighting of that dimension during perception and drive increased attentional gain to that dimension. Research found that, for example, Chinese L2 learners of English rely on pitch cues more than native English speakers when categorizing stress patterns in nonsense words (Wang, 2008). English and Mandarin Chinese native speakers used vowel quality as the primary and pitch as a secondary cue, while pitch was found to be completely disregarded by Russian speakers when identifying stress location in a disyllabic nonword "maba" that was phonologically and phonotactically permissible across these languages (Chrabaszcz, Winn, Lin, & Idsardi, 2014). Russian speakers relied more on duration and intensity cues instead. Mandarin Chinese speakers were also found to rely more on pitch when perceiving English speech prosody and musical beats (Jasmin, Sun, & Tierney, 2021) and overuse pitch when speaking relative to other auditory cues (Zhang, Nissen, & Francis, 2008). This over-reliance on pitch may slow L2 acquisition, contribute to having an identifiable non-native accent, and make one's prosody production harder to interpret, as it may inhibit extraction of linguistic structure from other useful cues such as amplitude, duration, and spectral shape (e.g., Escudero, Benders, & Lipski, 2009; Kong & Yoon, 2013; Kong & Edwards, 2015). Since, depending on their language background, listeners might weigh the available acoustic information differently, it is possible that the differences in dimension-selective attention or dimensional salience can be contributing factors to difficulties with not only L2 speech prosody acquisition, but speech perception and production more generally.

Despite their prominent role in cue weighting theories of speech perception (e.g., Francis & Nusbaum, 2002; Holt et al., 2018), attention to and salience of L1-relevant acoustic dimensions have not been thoroughly tested in the context of L2 speech acquisition. Some results pointed to the potential role of attentional shifts in language learning by showing changes in relative weights of F0 vs VOT depending on the listeners' L2 proficiency (English learners of Korean, Kong & Edwards, 2015; Korean learners of English, Kong,

& Kang, 2022), but no direct evidence of attentional enhancement was presented in these studies. In another study that investigated the role of attentional control in speech comprehension under challenging conditions, listeners with better non-verbal selective attention (operationalized as the accuracy of detected repetitions within the attended stream of tones) were better able to perceive speech in the presence of competing talkers (Tierney, Rosen, & Rosen, 2020). The role of selective attention was also implicated by results showing that people alter their strategies depending on the changes in cue usefulness (Idemaru & Holt, 2011). However, even stronger evidence comes from experiments testing the effects of informational masking. Findings demonstrated that perceptual strategies change when people are distracted either by the increased attentional load (completing arithmetical tasks, Gordon et al., 1993; Kong & Lee, 2018) or the presence of competing talkers (Symons, Holt, & Tierney, 2023), suggesting that listeners adaptively shift attention in response to demanding listening conditions. Recent research introduced a behavioural measure of dimension-selective attention (previously used in monitoring neural indices of attention with EEG, Symons, Dick, & Tierney, 2021) and showed that, along with other auditory processing measures (auditory acuity and auditory-motor integration), attention could predict phonological and morphosyntactic knowledge among Chinese learners of English with 0.5-10 years of immersion experience (Saito et al., under revision). A better ability to attend to various acoustic cues and flexibly adjust the focus of attention to integrate across available cues might be beneficial for successful L2 speech learning. However, the role of attention and dimensional salience in explaining individual differences in L2 speech perception and production is yet to be tested.

Present study

The goal of this study was twofold. Firstly, by building on the research in second language acquisition, I test the predictive power of language learning experience (i.e., age of acquisition, length of residence and daily L2 use) in explaining individual differences in L2 learning outcomes among learners with limited immersion experience. While the literature agrees that prolonged immersion in L2 environment leads to noticeable gains in L2 proficiency (Flege & Liu, 2001), it is less explored whether L2 learning experience can explain individual differences at the early stage of immersion in the L2 environment.

Moreover, past research on L2 acquisition has primarily focused on the segmental properties of speech (consonant production, Flege, Takagi, & Mann, 1995; vowel quality, Flege, MacKay, & Meador, 1999; Piske, Flege, MacKay, & Meador, 2002) or the degree of foreign accents (e.g., Flege, Yeni-Komshian, & Liu, 1999) and relatively little research has been devoted to acquisition of L2 prosody. Although scholars recognized the importance of prosody in L2 learning by showing how it relates to accentedness and comprehensibility (e.g., Munro & Derwing, 2001; Trofimovich & Baker, 2006) and emphasized the need for suprasegmental training[13], it is still unknown what factors support the acquisition of accurate L2 prosody perception and production. In this study, I attempt to fill that gap by examining the role of experiential, behavioural and neural factors in predicting L2 prosody perception and production. I intend to not only discern whether and what factors are helpful for developing L2 prosody, but also examine the role of these factors in predicting L2 proficiency across a range of other language measures, namely speech perception, vocabulary and grammar knowledge. I test whether and, if so, to what degree the same factors can explain L2 development across various aspects of language.

Secondly, inspired by the putative role of attention and dimensional salience in cue weighting theories, I test their contributions to explaining additional variance in L2 learning outcomes, once L2 learning experience variables are included. Since speech is the primary form of language input for most learners, auditory perceptual skills such as the ability to weight acoustic dimensions according to their reliability in predicting category

[13] It is worth mentioning that even though research recognized the importance of prosody in L2 speech perception, teaching prosody is not commonly included as a part of the L2 learning curriculum (Lengeris, 2012) or even showcased in the review of trends in L2 speech perception training (Ingvalson, Ettlinger, & Wong, 2014). It could be that because prosody concerns more nuanced differences at the level of words and phrases that might matter only in specific contexts, compared to individual phonemes' pronunciations which can change the meaning of each word, it is not of high priority. It is possible that in some cases the intended meaning can be deduced from the overall context of the conversation, but sometimes it might not be possible and could lead to miscommunication. This lack of attention to speech prosody is even more surprising given that very often native-like pronunciation is set as a goal for L2 learners and correct intonation is associated with the degree of perceived accentedness (e.g., Trofimovich & Baker, 2001).

membership (Francis, Baldwin, & Nusbaum, 2000; Toscano & McMurray, 2010) might serve as a bottleneck for language comprehension and pronunciation. Poorer speech comprehension could be a result of differential weighting across available acoustic cues or the inability to allocate attentional resources to relevant information. To test this hypothesis, I measure behavioural dimension-selective attention and dimensional salience. I use the EEG frequency tagging paradigm as a neural measure of salience in verbal and non-verbal sound sequences, where the degree of inter-trial phase coherence (ITPC) at the tagged dimension change rate represents the extent to which attention is being captured by that dimension. This paradigm allows me to examine the relative salience of acoustic dimensions while precisely controlling the informational content of these dimensions, but it does not reflect dimensional salience in real-life listening conditions. To provide a more ecologically valid measure of dimensional salience, I also measure neural entrainment to pitch and speech envelope during listening to continuous speech. The analysis uses the multivariate temporal response function (mTRF; Crosse et al., 2016, 2021) that estimates how well the selected stimulus features are represented in the brain. I predicted that both attention and salience would be important factors in predicting L2 speech acquisition and contribute to explaining differences in speech perception (vowel and prosody), but not necessarily in vocabulary knowledge, which may instead be linked to better working memory (Perez, 2020) or verbal memory (Atkins & Baddeley, 2008).

4.2 Methods

Participants

The study involved 60 native Mandarin speakers aged 18-31 (M=22.62, SD=3.27; 53 females, 6 males, 1 non-conforming)[14]. They were university students enrolled in BSc, MSc, and PhD programs, such as psychology, education, economics, law, architecture, or engineering. Participants were recruited from the SONA platform, social media groups, and societies for Chinese students in London (Facebook and WeChat). They spoke Mandarin as

[14] These are the same Mandarin speakers reported in Chapter 2 and 3.

their first language and learned English as a second language, starting from primary school ($M_{EFL\ age}$=7.58, $SD_{EFL\ age}$=3.05). Their experience with English language before coming to the UK was limited to the classroom context. They reported 1 to 17 months (M=7.41, SD=3.21) of L2 English immersion after arriving in the UK as young adults ($M_{AOA\ age}$=21.08, $SD_{AOA\ age}$=3.52). One participant visited US at the age of 11 for a few months, but they were not immersed in the L2 environment at that time. Seven participants reported speaking an additional foreign language (1 Russian, 1 French, 1 German, 2 Japanese, 2 Korean). Based on Zhang et al.'s (2020) criteria, 29 participants were musicians (≥6 years of musical training). Table 7 summarizes participants' language and musical experience.

Basic demographic information about participants' language and musical training was collected in the prescreening interview to confirm their eligibility for the study. To gather more detailed information about participants' language background, I used a custom questionnaire that assessed relevant variables for successful second language learning in naturalistic (Language Contact Profile: Freed, Dewey, Segalowitz, & Halter, 2004) and classroom settings (Foreign Language Experience Questionnaire: Saito & Hanzawa, 2016).

Behavioural measures

Dimension-selective attention tasks

Stimuli

The base stimuli were eight unique tokens, four verbal and four non-verbal. The verbal stimuli were two versions of the word "barbecue" (with and without emphasis) extracted from "Tom likes barbecue chicken" phrase downloaded from the Multidimensional Battery of Prosody Perception (Jasmin, Dick & Tierney, 2020). From both versions, I extracted the vowel /a/ and morphed them along fundamental frequency (F0) using STRAIGHT voice morphing software (Kawahara & Irino, 2005). I created a continuum of 100 samples, where only pitch varied and other acoustic parameters remained constant. Two stimuli were selected, level 1 sample (110.88 Hz) and level 56 approximately 2 semitones higher than level 1 (124.40 Hz). These values were chosen to guarantee that differences across both dimensions will be easily perceivable. Then, base stimuli were manipulated in length (70.58 ms and 175.83 ms) using Praat software (Boersma & Weenink, 2001). The resulting stimulus grid consisted of a 2 (pitch) x 2 (duration) configuration with four distinct tokens. Non-verbal stimuli were generated in MATLAB to match the acoustic properties of the verbal sounds. They were four unique complex tones with 4 harmonics and varied along pitch and duration (110.91–124.72 Hz and 70–175 ms).

The base verbal and non-verbal stimuli were combined into 2 Hz sequences where pitch and duration varied simultaneously but at different rates (0.67 Hz and 1 Hz). In half of the sequences for each dimension, I introduced repetitions which resulted in four types of trials: pitch repetition only, duration repetition only, repetitions in both dimensions and no repetitions in either dimension. The stimuli used in each trial type and attention condition were identical, with the only difference being the focus of attention. From each set of stimuli, 64 were randomly chosen and assigned to either the "attend to pitch" or "attend to duration" conditions, with 32 trials per attention condition. In two versions of the task (versions A and B), the stimuli were assigned to opposite attention conditions. For example, if a recording was assigned to the "attend to pitch" condition in version A, it would be assigned to the "attend to duration" condition in version B. This ensured that the stimuli were counterbalanced across attention conditions in both task versions.

Task

During the experiment, participants listened to sequences of verbal and non-verbal sounds varying in pitch and duration at two rates. At the beginning of each block, participants were instructed to focus their attention on changes in one of these dimensions. Each trial began with a 500 ms long silence, followed by the sound sequence and a prompt on the screen asking participants whether they detected a repetition within the attended dimension. Participants indicated their responses by clicking the 'Yes' or 'No' button on the screen and were provided with feedback on a trial-by-trial basis. Short breaks were included at the end of each block. Prior to the main task, participants familiarized themselves with the different pitch and duration levels by listening to example stimuli with and without repetitions. Subsequently, they completed a brief training task closely resembling the main task, with the difference that only the attended dimension varied. To proceed to the main task, participants needed to achieve a minimum of 75% accuracy (6 out of 8 trials), and up to 3 attempts were permitted. Participants who did not meet the performance threshold were not allowed to continue with the study. The task consisted of 4 blocks for each type of stimulus (verbal and non-verbal), corresponding to 2 attention conditions and 2 rates of change. At the start of each block, participants were informed about the specific dimension to attend to and the expected rate of change. The final score was calculated by combining the hit rates across the dimension change rates for each attended dimension.

Language assessments

Prosody perception test

The prosody perception test is a custom-built task designed to measure participants' ability to distinguish different prosodic features (i.e., linguistic focus, phrase boundary and lexical stress) that are important to deliver intended meaning to the listeners during spoken communication.

Stimuli

The lists of phrase boundary and linguistic focus stimuli for the test were taken from the Multidimensional Battery of Prosody Perception (MBOPP, Jasmin et al., 2020) dataset.
I also included additional sentences selected for lexical stress to capture the contrasts between all prosodic features (i.e., two sentences per contrast; Table 11). 20 contrasts for each feature were chosen for the test (i.e., 40 sentences per feature, 120 sentences in total; see Supplementary Materials A for full stimuli list). This was to guarantee a good representation of each feature in the stimuli set and keep the time needed to complete the task relatively short (<15 minutes to complete). To not expose participants to the same voices they would be listening to during the cue weighting categorization task, I recorded the selected stimuli with two new voice actors (1 female, 1 male; both native British English speakers). All the stimuli were recorded in a sound-proof booth at Birkbeck University with a RØDE NT1A large-diaphragm cardioid condenser microphone with shock mount and pop filter and Audacity software (version 3.0.5). Voice actors were instructed by the researcher to read the sentences aloud using their usual speaking strategies to clearly convey the linguistic contrasts while maintaining the natural tone of their voice.

Table 11. Examples of prosody perception task stimuli across prosodic features. Capitalization indicates contrastive stress and emphasis.

Prosodic feature	**Token A**	**Token B**
Linguistic focus	***Early focus***	***Late focus***
	FLYING planes	flying PLANES
	PLANT flowers	plant FLOWERS
Phrase boundary	***Early closure***	***Late closure***
	When Mary helps, the homeless	When Mary helps the homeless
	When Paul drinks, the rum	When Paul drinks the rum
Lexical stress	***1st syllable stress***	***2nd syllable stress***
	PROtest	proTEST
	CONtent	conTENT

Two versions of each phrase (Token A and Token B) were then morphed with the STRAIGHT morphing toolbox for MATLAB (Kawahara & Irino, 2005). The morphing procedure involved time aligning the stimuli based on the visual inspection of similarity matrices and manually marking anchor points for morphing (i.e., points corresponding to the same contents in both recordings). After marking those points, I created the stimulus continua by manipulating the levels of fundamental frequency (F0) and duration with 5% increments. Stimuli levels of 70 vs 30% were chosen for inclusion in the prosody test across all prosody features based on the pilot experiment[15]. Selected stimuli were

[15] The pilot experiment was conducted online with the prosody test designed and hosted on Gorilla. 16 native Mandarin Chinese speakers recruited from Prolific recruitment platform took part. They were all adult L2 English learners. First, I tested the stimuli with 20 vs 80% levels of pitch of duration. 8 participants (4 females, M_{age}=28, SD_{age}=5.07, 5 subjects reported up to 5 years of immersion experience) were tested. Their performance reached ceiling (% correct responses: M_{focus}=81.56, SD_{focus}=12.39, $range_{focus}$=60-97.5, M_{phrase}=91.56, SD_{phrase}=6.26, $range_{phrase}$=80-100, M_{stress}=80.94, SD_{stress}=14.51, $range_{stress}$=52.5-95). I increased the perceivable difficulty of the task by selecting the 30 vs 70% stimuli for listening in the second listening experiment. 8 different participants (7 females, M_{age}=28.75, SD_{age}=6.48, 4 reported up to 5 years of immersion experience) completed the second version of the task. Their performance did not reach ceiling and was more varied across participants (M_{focus}=80.31, SD_{focus}=10.30, $range_{focus}$=65-92.5, M_{phrase}=77.5, SD_{phrase}=6.68, $range_{phrase}$=65-85, M_{stress}=71.25, SD_{stress}=13.69, $range_{stress}$=50-92.5). I concluded that this level of stimuli would allow me to observe individual differences in prosody perception performance.

randomly assigned to one of the two versions of the task. To counterbalance the presence of male and female voices in both versions, the stimuli were assigned to the opposite conditions. For example, the focus05_pitch30_dur30 recording with a male voice and focus05_pitch70_dur70 with a female voice were assigned to one task version. In the second version, the selected focus05_pitch30_dur30 token would have a female voice and focus05_pitch70_dur70 male voice. There were 120 trials organized in 3 blocks by linguistic feature in each version. Presentation of blocks and trials within blocks was randomized across participants.

Procedure

On each trial, participants heard a spoken phrase or a word and saw two versions of it displayed on the screen. Their task was to decide whether the spoken phrase or word sounded most like the phrase or word written on the left or the right side of the screen. I asked participants to press the button to indicate their responses as quickly as possible without making mistakes. I asked them to be quick because I was interested in participants' spontaneous processing of language and wanted to control for participants' careful monitoring of their performance using explicit L2 knowledge (Saito & Plonsky, 2019). The final score was calculated as the proportion of correct responses per feature.

Segmental speech perception test

A speech perception test was used to assess participants' ability to perceive English minimal pairs that were recognized as being difficult to acquire for Mandarin Chinese learners of the English language due to cross-linguistic differences between English and Mandarin Chinese phonetic systems (e.g., Ruan & Saito, 2023). Stimuli consisted of 42 words forming 21 minimal pairs selected to represent the phonological contrast between the long [ɪː] and the short [ɪ] vowels (e.g., sheep vs ship), which are particularly difficult for Mandarin Chinese speakers to learn. Most Chinese learners cannot distinguish that contrast in English speech and have trouble pronouncing it accurately. The vowels can be differentiated not only by the mouth position (i.e., the vowel [ɪ] is pronounced lower than [ɪː] and differs in formant frequencies; Table 12) but also by their relative length. Although the duration is not the primary cue for distinguishing the two, the [ɪː] vowel was found to be consistently and reliably longer for most L2 (e.g., Hillebrand, Getty, Clark, & Wheeler, 1995) and Received Pronunciation speakers (Ladefoged & Johnson, 2014).

Table 12. Example formant frequencies for adult male (Wells, 1962). F1 and F2 vary more than F3. Perception of English [i] and [ɪ] contrast is defined by the combination of temporal and spectral cues (longer, lower F1, and higher F2 for [i] and shorter, higher F1, and lower F2; (Hillenbrand, Getty, Clark, & Wheeler, 1995).

Vowel	***F1 (Hz)***	***F2 (Hz)***	***F3 (Hz)***	***Character in SSBE***
[ɪː]	280	2620	3380	Close, front, long
[ɪ]	360	2220	2960	Close, front, short

All stimuli for the speech perception test were recorded by one male and one female speaker of Standard Southern British English (SSBE), resulting in 84 unique trials (2 speakers, 21 minimal pairs). Participants were asked to listen to a spoken word and then to indicate the correct spelling from the two options displayed on the screen by pressing the button with the appropriate word written on it. They heard each word only once and could not repeat it. The outcome measure was a portion of correct responses.

Grammatical judgement task

A timed version of the Grammatical Judgment Test (GJT; Godfroid, Loewen, Jung, & Park, 2015) was used to measure participants' ability to indicate the relative acceptability of errors in written sentences. With this test, I measured the implicit grammatical knowledge that allows people to make fast and intuitive judgements about the correctness of sentences. As such, participants were not required to be aware of the type of error or know the correct solution, they just needed to recognize whether the sentences were grammatically correct or incorrect. Such a procedure is meant to tap into their procedural representations of L2 English grammar, rather than declarative knowledge of grammatical rules. The test included grammatical and ungrammatical versions of 68 English sentences spanning 17 linguistic features (i.e., 136 sentences in total). A multitude of examples was included to gain a reliable measure of grammatical acceptability of all grammatical structures in the test (i.e., 8 sentences per grammatical structure). The grammatical forms included a wide selection of grammar aspects that are difficult for L2 speakers to learn (e.g., plurals -s, possessives, indefinite articles, past tense). The final score was calculated as a sum of correct identifications of grammatical sentences and correct rejections of ungrammatical sentences. Participants were presented with a series of sentences, one sentence at a time, and asked to indicate their grammatical acceptability by pressing the "grammatical" or "ungrammatical" buttons on the screen. Participants were given only a few seconds to read each sentence and respond to prevent them from inspecting each item and overthinking its structure. The time limit varied from item to item, depending on stimulus length, following the time limits established by Godfroid et al. (2015).

Vocabulary knowledge test

To quickly measure participants' general English proficiency, I used LexTALE (Lemhoefer & Boersma, 2012). This 5-minute vocabulary test was shown to correlate highly with standard language proficiency measures such as the Global Scale of Common European Framework of Reference (CEFR) or levels from the Quick Placement Test (2001). The test involved a series of brief lexical decisions measuring the vocabulary proficiency of non-native speakers. Participants saw one word at a time and needed to decide for each word whether it was an existing English word or a non-word by pressing the "Yes" or "No" buttons on the screen. The participants were instructed to select “Yes” for the words they know and use or recognize even if they did not remember their meaning. If they were unsure if the word was a real English word, they were meant to select "No". Stimuli were 40 English words and 20 non-words, all between 4 and 12 letters. The scores were averaged proportions of correct responses (%). I corrected for the unequal number of word and non-word items by averaging the percentages for these two item types to prevent the high error rate bias for the less numerous non-word items.

Picture description task

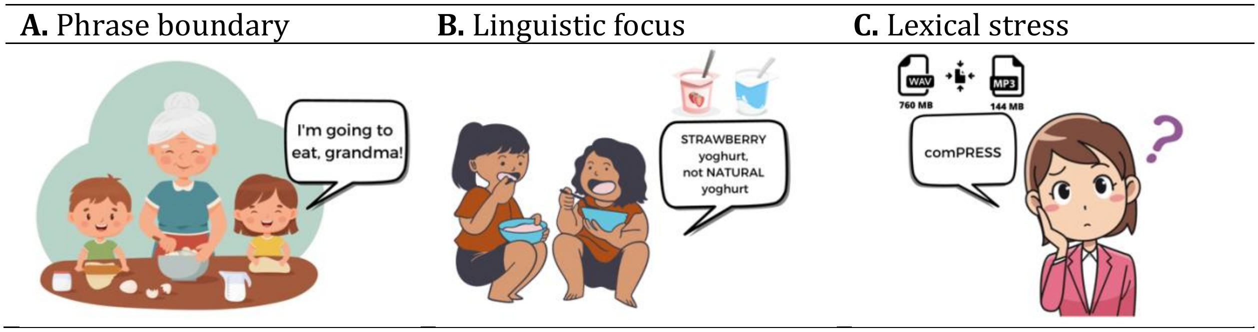
I adapted the classic picture description task widely used in speech production research (e.g., Saito & Hanzawa, 2016) to measure participants' spontaneous production of English prosody. I designed pictures including prompts comprising phrases reflective of the prosodic contrasts of interest: phrase boundary, linguistic focus, and lexical stress.

Stimuli

The stimuli were words, phrases or sentence pairs capturing the lexical stress, linguistic focus and phrase boundary differences embedded into colorful cartoon images. To avoid familiarity effects the sentences included in the production task differed from prosody contrasts included in cue weighting tasks and prosody test. The images were designed to correspond thematically with the embedded text and were created with the Canva platform (*www.canva.com*) and its open library of graphics. I created 24 unique images in total, 4 contrasts per feature (for full list of stimuli see Supplementary Materials B). The images were split into two sets, Version A and B, randomized between testing sessions to avoid training effects and familiarity with picture content.

Procedure

Participants were asked to describe each picture by using the target phrase or word embedded within a picture. They saw one picture at a time, and they had 10 seconds to prepare (i.e., look at the picture and think about what to say) and then 30 seconds to speak. They saw a timer with a countdown indicating the remaining time on the screen. Before moving on to the task, the researcher presented participants with three examples of picture descriptions to indicate what kind of descriptions they should provide (Figure 15).



A. Phrase boundary **B.** Linguistic focus **C.** Lexical stress

I'm going to eat, grandma!
STRAWBERRY yoghurt, not NATURAL yoghurt
WAV
760 MB
MP3
144 MB
comPRESS

Figure 15. Examples from prosody picture description task instruction. During the testing session, the researcher demonstrated these images and performed sample descriptions as examples. (A) Phrase boundary example: "In this picture, I can see an old lady and two kids. I think they are baking something. At some point, the girl, Jenny, says, "I'm going to eat, grandma!". Most probably, she's hungry. They all look very happy." (B) Linguistic focus example: "Two girls in this picture are eating something from their bowls. I think it's a strawberry yoghurt, not a natural yoghurt, because what they have in their bowls is pink. Pink yoghurt is more likely strawberry. I don't like strawberry yoghurt." (C) Lexical stress example: "There is a lady in a pink suit who looks troubled. She doesn't know how to compress wav files into mp3. Her boss will be angry if she doesn't do it". If participants didn't know the meaning of the word, they were advised to attempt to describe the picture by including that word in their description, for example, by saying: "There is a lady that is thinking about something. She says "compress", so she most probably needs to compress something, but she doesn't know what to do" or even "The picture shows a working woman. I don't know what the word "compress" means, but the lady is saying it. She's wearing a pink suit and a nice haircut." This was to encourage the articulation of all the prosodic contrasts even if participants were not familiar with the target words and their meaning.

Speech recordings

Participants' responses were recorded with RØDE NT1A large-diaphragm cardioid condenser microphone with shock mount and pop filter and Audacity software (version 3.0.5). For safety reasons during the COVID-19 pandemic[16], participants were wearing face masks during the testing sessions. Whenever possible, participants were allowed to remove their masks for the duration of the speaking task (less than 10 minutes) and put their masks on for the remainder of the session. For participants who were uncomfortable

[16] Although COVID-19 restrictions are no longer present, the data collection for this project began during the Summer of 2021, when such policies were still in place. The restrictions were lifted halfway through the project (data collection continued until the Autumn of 2022), so to keep the experimental conditions constant across all participants and testing sessions, I followed the same protocol.

removing their masks in the lab, I provided standard two-layered surgical masks to keep the speech distortions constant across participants and testing sessions. Surgical masks of this type were shown to cause minor speech distortions in controlled speech recordings (Corey, Jones, & Singer, 2020; Magee et al., 2020; Nguyen et al., 2021).

Speech pre-processing

First, all the recordings[17] were manually trimmed to contain only 30 seconds corresponding to each picture participants described, so any comments from the researcher and between-trials discussions were removed. Then, from each recording, the researcher manually extracted the excerpts that captured the prosodic contrasts (identical portions across the recordings).

Sometimes participants did not complete the sentence from the prompt or expanded the wording. For example, they only said, "GREY cat" instead of using the full contrastive structure of "GREY cat, not BROWN cat" or added description between the first part of the contrast and the second (e.g., "The girl likes FRENCH food which is tasty and healthy, not ITALIAN food" instead of "FRENCH food, not ITALIAN food"). I accepted such incomplete or interrupted contrasts, but only if these changes did not interfere with the overall prosody of the contrast. I excluded examples such as “study theory of MUSIC” instead of “study MUSIC” because including the additional words changed the structure of the phrase. Participants were aware of contrastive focus, and so, I expected them to pronounce the target phrases accordingly regardless of their exact lexical context in which they were presented.

I also found several instances of modified contrasts where participants used the target word from the prompt but changed its form (this happened for pronunciations of lexical stress). For example, participants said "refusing" instead of "refuse", and "presenting" instead of "present". In these cases, I extracted only the root of the verb without the “ing”

[17] There were 1440 recordings in total recorded from 60 participants completing the task twice (pre-test and post-test session) with 12 examples. Here, only the data from the first session is reported.

ending. I accepted suffixes such as "s" added to "present" (resulting in "presents") because they could be plural ("two Christmas presents") or verb forms ("she presents"). But I did not accept instances such as "presentation" (instead of a verb "present") or "projector" (instead of a verb "project") as they are nouns. I kept minor word additions (e.g., "but", "and") to the target phrases if cutting them would introduce unwanted distortions. For example, sometimes I kept "and when the boy leaves the house" instead of "when the boy leaves the house", which did not interfere with the overall prosody of the phrase. Replacement of single words (e.g., "he" instead of "the boy" in "when the boy leaves the house) was also acceptable. I also left extra wording or filled pauses if they were mixed with target words in a manner which did not change the overall prosody of the utterance. For example, I kept filled pauses in sentences like "when he uhm leaves the house" or extra words like "when the little boy visits his grandpa" (instead of "when the boy visits his grandpa"). In case of repeated or corrected target words, I included only the first exemplar, regardless of whether it was pronounced correctly or incorrectly.

Selected samples were manually inspected for background noise and audio quality. To ensure comparable quality across recordings for the rating session, background noise was removed with the "DeNoise" function (stimuli waveforms and spectrograms were visually and aurally inspected to minimize distortions)[18] and normalized to -3 dB with the "Normalize process" effect in Adobe Audition (2021) software.

Speech production ratings

Raters

All speech samples were rated by 5 native Southern British English speakers (all females, M_{age}=26, SD_{age}=2.74, $range_{age}$=23-28). All raters had previous experience conducting L2 speech ratings and were highly familiar with variability of native English prosody articulation and variability of non-native accents. Three participants were current doctoral

[18] The DeNoise function uses state of the art machine learning algorithms to remove background noise from audio files. I used the light noise reduction setting with processing focus across all frequencies. I adjusted noise reduction to only 15% across files to prevent any destructive interference with participants' speech.

students (including one who obtained some training in L2 linguistics), one a BA graduate in English literature and one a working professional. Although they were all naïve listeners, the task involved rating aspects of speech every native speaker should be familiar with and their familiarity with these constructs was confirmed by their self-rated understanding of the task (ratings on a 9-point scale where 9 represents very good understanding of the task; M_{focus}=8.6; M_{stress}=8.4, M_{phrase}=8.2).

Rating procedure

The rating session comprised three main blocks representing three prosodic features of interest: linguistic focus, phrase boundary and lexical stress. Listeners were told they would be rating speech produced by L2 learners and native speakers. First, the raters were introduced to linguistic concepts they would be rating by presenting them with short definitions and the corresponding audio examples recorded by native speakers of English. Detailed instructions asking the raters to focus on those aspects while ignoring other pronunciation errors at the suprasegmental level (e.g., mispronouncing vowels or consonants, foreign accent, speech impediment) were displayed before the beginning of each block. Each trial started by playing the speech sample for rating and then the scale appeared on the screen. A repeat button allowed the raters to listen to problematic samples again (up to 3 times). Raters could also leave notes in the text box to mark any trials where the audio was not playing properly (1 trial was removed from analysis due to issues with audio playback) and make comments (e.g., one rater reported an issue with screen resolution that was obscuring part of the scale so they typed their ratings for several trials in the comment window instead and these were manually added to the dataset). Raters completed their task in 3 sessions, approximately 1-1.5 hours long each. The order of blocks across participants and trials within each block were randomized.

Rating scales

I adopted the 9-point rating procedure developed by Saito (2013; the procedure originally created to capture the different developmental stages of English /r/ vs /l/ acquisition by Japanese raters) to measure how well participants can articulate L2 English prosody. Although the focus of this study is on prosody rather than on specific segmental contrasts exemplified in the original procedure, the basic assumptions about the distinct phases of the L2 production development are still applicable here. All listeners judged the quality of prosodic contrasts by choosing one of the response alternatives presented in Table 13. The ratings were meant to represent the accuracy of participants' prosody production. For example, for 1st syllable stress lower ratings related to more native-like productions, whereas for 2nd syllable stress higher ratings meant better productions. The mid-point of the scale was used for neutral samples that could reflect the interlanguage stage of adult L2 learning (for similar approach to acquisition of segmentals see Saito et al., 2022).

Table 13. Rating scales descriptors. The scale and its descriptors were adapted from Saito (2013ab) to capture the different stages of L2 prosody acquisition.

Rating	**Phrase boundary** ***(comma placement)***	**Linguistic focus** ***(word emphasis)***	**Lexical stress** ***(syllable emphasis)***
1	Nativelike early boundary	Nativelike emphasis on the 1st word	Nativelike stress on the 1st syllable
2	Good early boundary	Good emphasis on the 1st word	Good stress on the 1st syllable
3	Probably early boundary	Probably emphasis on the 1st word	Probably stress on the 1st syllable
4	Possibly early boundary	Possibly emphasis on the 1st word	Possibly stress on the 1st syllable
5	Neutral exemplars	Neutral exemplars	Neutral exemplars
6	Possibly late boundary	Possibly emphasis on the 2nd word	Possibly stress on the 2nd syllable
7	Probably late boundary	Probably emphasis on the 2nd word	Probably stress on the 2nd syllable
8	Good late boundary	Good emphasis on the 2nd word	Good stress on the 2nd syllable
9	Nativelike late boundary	Nativelike emphasis on the 2nd word	Nativelike stress on the 2nd syllable

Ratings reliability and final scores

To assess the quality of ratings, I computed Pearson's correlations between all raters for each linguistic feature separately. Based on these correlations, I selected ratings of Raters 1, 2 and 3 for further analyses who showed the best correlations with one another (Table 14). Then, to test the agreement across selected raters, I computed the inter-rater correlation coefficient (ICC) with a two-way random-effects model (Shrout & Fleiss, 1979) implemented within the ICC function from psych R package. Agreement between raters was moderate (ICC$_{\text{stress}}$=.60; ICC$_{\text{phrase}}$=.68; ICC$_{\text{focus}}$=.52; interpretation according to Koo & Li, 2016). The reason for relatively low inter-rater correlations might be twofold. First, participants' realizations of prosodic features varied significantly in their correctness and way of articulation and were inconsistent across items (i.e., production-related variability). Second, despite obtaining identical instructions, it is possible that the raters focused on different aspects of prosody. For example, they could give more or less weight in their judgements to how prosodic contrasts were articulated (i.e., individual differences in raters' cue weighting) or be more or less sensitive to foreign-accented articulation. They were explicitly asked to ignore aspects of pronunciation not related to prosody, such as for example, coarticulation effects, misspellings, and heavy accent, but it is likely that these aspects affected their ratings in some way due to their strong influence on prosody.

Table 14. Correlation matrix representing Pearson correlation coefficients between the 9-point ratings of lexical stress, phrase boundary and linguistic focus by 5 raters.

Lexical stress		**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**
	Rater 1	1.00				
	Rater 2	.59	1.00			
	Rater 3	.55	.73	1.00		
	Rater 4	.42	.50	.42	1.00	
	Rater 5	.35	.42	.31	.29	1.00
Phrase boundary		**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**
	Rater 1	1.00				
	Rater 2	.70	1.00			
	Rater 3	.66	.75	1.00		
	Rater 4	.58	.66	.67	1.00	
	Rater 5	.66	.70	.67	.65	1.00
Linguistic focus		**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**
	Rater 1	1.00				
	Rater 2	.64	1.00			
	Rater 3	.52	.55	1.00		
	Rater 4	.51	.59	.45	1.00	
	Rater 5	.33	.43	.33	.33	1.00

To compute the final production accuracy scores, the lower end of the scale (i.e., 1-5 ratings) was inverted using the 10-N formula so that the native-like productions of both categories were equal to 9 (i.e., the higher the rating, the more native-like their production). Then, I dichotomized the ratings into correct vs incorrect realizations of the prosodic contrasts. To do so, I counted all the trials with ratings above 5 as correct (representing a range of correct pronunciations) and all the trials with ratings equal to or lower than 5 as incorrect (i.e., receiving a neutral rating or representing the opposite end of the contrast). I obtained the final score for each participant by averaging all correct and incorrect responses across all trials and raters. That procedure resulted in a single number per participant per feature (the higher the value, the higher the accuracy of prosody production).

Neural measures

EEG data acquisition

All EEG data was recorded from 32 Ag-Cl active electrodes using a Biosemi™ ActiveTwo system with the 10/20 electrode montage. Data were recorded at a 16,384 Hz sampling rate and digitized with a 24-bit resolution. Two external reference electrodes were placed on both earlobes for off-line re-referencing. Impedance was kept below 20 kΩ throughout the testing session. Triggers marking the beginning of each trial (every 6 tones, or 1.2 seconds for the frequency tagging paradigm) or each block (for the dimensional salience during naturalistic speech perception paradigm) were recorded from trigger pulses and sent to the data collection computer. All EEG data processing and analysis were carried out in MATLAB (MathWorks, Inc) using the FieldTrip M/EEG analysis toolbox (Oostenveld et al., 2011) in combination with in-house scripts.

Frequency tagging paradigm

Stimuli

The base stimuli were speech and tones sampling a 2 (pitch) x 2 (duration) space created for the dimension-selective attention task described in the earlier section. They were combined into 5 Hz sequences where pitch and duration changed at different rates (1.67 Hz or 2.5 Hz). To keep these changes unpredictable, 20 repetitions were inserted into each sequence. Additionally, a subset of 3-5 stimuli per sequence (32 in total) had their amplitudes reduced by 25% (-12.04 dB) to function as amplitude oddballs. The timing of oddballs was randomized in each sequence, but they could not occur within the initial or final 4.8 seconds (4 epochs) and were separated by a minimum of 4.8 seconds from one another. All participants were presented with the same set of sequences, and the order of presentation was counterbalanced. The stimuli were presented diotically using ER-3A insert earphones (Etymotic Research, Elk Grove Village, IL) at a maximum sound pressure level of 80 dB SPL, with a sampling rate of 44100 Hz. PsychoPy3 software (version 3.2.3) was used for stimulus presentation.

Behavioural task

Participants listened to sound sequences and responded by pressing keys when they heard occasional quiet tones. This task was meant to keep participants focused on the sound without indicating any specific acoustic characteristics. Before the EEG recording began, participants completed a brief practice session. They listened to verbal and non-verbal sound sequences for approximately one minute each and had to reach a minimum of 5 out of 6 correct responses. Visual feedback was provided during the practice, showing the number of correct and incorrect responses, and missed targets. Most participants completed the practice on their first attempt, while others were asked to repeat it to ensure optimal performance during EEG recording. The main task was identical to the practice run but included longer sound sequences (4 blocks, each containing four 2-minute sequences) and had no visual feedback.

Intertrial phase coherence (ITPC)

The recorded data were resampled to a frequency of 512 Hz and referenced to the average of the earlobe electrodes. Subsequently, a low-pass zero-phase sixth-order 30 Hz Butterworth filter and a high-pass fourth-order zero-phase 0.5 Hz Butterworth filter were applied. The data were then divided into epochs of 1.2 seconds based on the recorded triggers. Independent component analysis (ICA) was performed to correct for eye blinks and horizontal eye movements, and components associated with these artifacts were visually identified and removed. Next, the data from the 9 channels with the strongest ITPC (i.e., AF4, F3, Fz, F4, FC1, FC2, FC5, Cz, and C3) across all participants (N=60) was extracted, forming a cluster of frontocentral channels. The data were averaged across these selected channels. Any remaining artifacts exceeding +/- 100 μV were discarded. To evaluate the cortical tracking of acoustic dimensions, inter-trial phase coherence (ITPC) was computed at the frequencies corresponding to the rate of dimension change. This involved applying a Hanning-windowed fast Fourier transform to each 1.2-second epoch, converting the complex vector at each frequency to a unit vector, and averaging across trials. The resulting length of the average vector served as a measure of phase consistency, ranging from 0 (indicating no phase consistency) to 1 (representing perfect phase consistency). The magnitude of ITPC at the frequency associated with a specific dimension served as an indicator of the salience of that dimension.

Dimensional salience during naturalistic speech perception

To examine dimensional salience during naturalistic speech perception, I recorded participants' brain activity while listening to excerpts from an audiobook. I measured neural entrainment to pitch contour and amplitude envelope of continuous speech. I analyzed the data using the mTRF toolbox for MATLAB (Crosse et al., 2016, 2021), which allows tracking neural encoding of pitch and amplitude in speech. The analysis relies on a temporal response function (TRF) to build models predicting neural responses from observed changes in the speech signal, which were shown to be reliable predictors of neural response to speech (e.g., Broderick & Lalor, 2020; Broderick et al., 2018, 2021).

Stimuli

Stimuli were the excerpts from the audiobook version of a classic novel "Old man and the sea" by Ernest Hemingway. These excerpts were successfully used in previous studies to measure encoding of various acoustic (e.g., amplitude envelope, Drennan & Lalor, 2019; amplitude envelope, relative pitch and resolvability, Teoh, Cappelloni, & Lalor, 2019) and linguistic features (e.g., semantic dissimilarity and lexical surprisal, Broderick et al., 2021) in the brain. The audiobook was read by a professional American English male narrator. The average speech rate was 210 words/min. Participants listened to 10 approximately 180 s long trials. To preserve the storyline, trials were presented chronologically without any repetitions or discontinuities.

Procedure

During the EEG recording, subjects were asked to focus on the narrator's voice and listen to the story. Stimuli were presented diotically at max 80 dB SPL at sampling rate of 44100 Hz using PsychoPy3 (v 3.2.3) via insert earphones ER-3A (Etymotic research, Elk Grove Village IL). There was no behavioural task. Participants were asked to relax and listen to the story while keeping their eyes open and restraining from any unnecessary bodily movements. During each session (pre-test and post-test) participants listened to two audiobook fragments (approximately 12-15 minutes each) with a short break in between the excerpts.

Data pre-processing

The neural signal was down sampled to 128 Hz and all channels re-referenced to the average of the reference electrodes placed on both earlobes. EEG data were then filtered with 0.2 Hz high-pass and 30 Hz low-pass 4th order Butterworth filters. To remove eye blinks, I performed the Independent Component Analysis (ICA) decomposition of psychophysiological data using the “runica” function from FieldTrip toolbox (infomax ICA algorithm of Bell & Sejnowski (1995) with the natural gradient feature of Amari, Cichocki & Yang, 1995, and the extended-ICA algorithm of Lee, Girolami & Sejnowski, 1999). Data processing was performed in MATLAB (The MathWorks Inc., 2020) using the combination of custom script with FieldTrip toolbox (Oostenveld et al., 2011).

Stimuli representations

In this study, speech stimuli are represented in terms of two parameters – speech envelope and relative pitch (both acoustic vector representations previously used by Teoh, Cappelloni, & Lalor, 2019) and linked the recorded neural signal to changes in these properties of speech stimuli. Speech envelope (time x amplitude) is one of the most widely used stimulus features used to represent continuous speech (e.g., Drennan & Lalor, 2019). Speech envelope was calculated by taking the absolute value of its Hilbert transform and then resampled to 128 Hz using the decimate function in MATLAB (Drennan & Lalor, 2019). To obtain a speaker-normalized relative pitch, a continuous measure of pitch (fundamental frequency/absolute pitch) was extracted using the autocorrelation method

(Boersma, 1993) in Praat software (Boersma & Weenink, 2016) at a sampling rate of 128 Hz and then z-scored.

Temporal response function (TRF) estimation

I estimated TRFs separately for speech envelope and relative pitch using the mTRF Toolbox (Crosse et al., 2016; 2021). The linear mapping between these features and EEG signal at each channel was described with the following model:

$$r(t,n) = \sum_{r} w(\tau,n)s(t-\tau) + \varepsilon(t,n)$$

Where $r(t,n)$ is the neural response sampled at time t and channels n consisting of the convolution of a particular stimulus feature vector $s(t-\tau)$ at various time lags τ with an unknown channel specific system response (i.e., TRF) $w(\tau,n)$, and $\varepsilon(t,n)$ represents the residual error not explained by the model. The range of time lags relative to the onset of each feature included the values typically used to capture the cortical response components of the evoked response potentials (ERP) and an additional 50 ms at each end to account for any regression artefacts (i.e., time window of -150 to 450 ms). Prior to analysis, 500 ms at the beginning of each excerpt was removed to avoid EEG onset artefacts.

The TRF weights were computed using ridge regression with the regularization parameter λ controlling the weighting of the penalty to the loss function (i.e., λ is a tuning parameter that balances between optimal data fit and bias). The optimal λ values were determined with the leave one out cross-validation using the mTRFcrossval function (Crosse et al., 2016). For each participant, averaged TRFs for each channel were calculated for every ridge parameter (λ = [2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^10]) for each trial. Then, the TRF is estimated by minimizing the mean-squared errors (MSEs) between the actual neural response $r(t,n)$ and predicted EEG response $\hat{r}(t,n)$:

$$\min \varepsilon(t,n) = \sum_{t} [r(t,n) - \hat{r}(t,n)]^2$$

This is solved by averaging the MSEs over trials, channels and participants and the optimal ridge parameter that showed the lowest MSE was selected is used to train the model.

Then, I computed the Pearson's *r* correlation coefficient between the actual brain response and predicted EEG responses across participants. The r values were averaged across trials and all EEG channels to prevent overfitting and optimize the model without selecting specific electrode locations. Prediction accuracy for each feature (speech envelope and relative pitch) was used as a measure of encoding strength or, in other words, of how well the feature is encoded in the brain.

General procedure

Interested participants who responded to the advertisements were invited to attend a brief telephone or video call with the researcher, during which the researcher asked questions about participants' basic demographics, language skills, and musical background to confirm candidates' eligibility to join the study. The researcher also explained the experimental procedure and task instructions and addressed participants' questions. After giving informed consent, eligible participants obtained links to online tasks (i.e., behavioural and language assessments excluding speech production task). All the online tasks were designed and hosted on the Gorilla Experiment Builder platform (Anwyl-Irvine et al., 2020). Upon the completion of online tasks, participants were invited to the lab for the EEG recording and speech production assessment[19]. The data analyzed in this chapter includes the behavioural and neural data from native Mandarin speakers at Time 1 (i.e., at Pre-Test) as outlined in Figure 1. The data collection took place at the Department of Psychological Sciences at Birkbeck, University of London, and adhered to all approved ethics procedures set by the departmental Ethics Committee. To compensate for their time, all participants received either cash (£10 per hour) or the equivalent amount in course credits. Processed

[19] After attending the EEG session, participants completed several days of online training and post-training online assessments and returned to the lab for a second EEG session. Data collected during these stages will be discussed in detail in Chapter 5.

data, analysis scripts and Supplementary Materials can be found at: *https://osf.io/5j8pe/?view_only=874b45bef48c4839807867f12a02809a*.

4.3 Results

Summary of performance across language assessments

I employed a comprehensive assessment battery consisting of five tests to evaluate various aspects of language proficiency. These tests covered measures of speech perception, prosody perception and production, vocabulary and grammar knowledge. The speech perception test assessed participants' ability to perceive L2 vowels accurately. The prosody perception test focused on evaluating the accuracy of stress and intonation patterns detection, whereas the prosody production task involved examining participants' ability to produce and manipulate prosodic features in their own speech. The vocabulary knowledge test aimed to gauge participants' lexical repertoire, and lastly, the grammatical judgement task evaluated participants' automatized knowledge of L2 grammatical structures. Summary measures for all language assessments are presented in Table 15.

Table 15. Summary of L2 assessments scores and their test-retest reliability

	M (SD)	***Range***	***Reliability***
Vocabulary (LexTale)	69.33 (12.03)	41.25 – 97.50	.746*
Words	75.58 (15.84)	27.5 – 100	.612*
Non-words	63.08 (24.70)	5 – 100	.749*
Grammar (GJT)	64.98 (11.58)	36.02 – 86.76	.605*
Grammatical items	81.15 (11.00)	52.94 – 95.59	.516*
Ungrammatical items	48.80 (16.99)	13.23 – 80.88	.740*
Speech perception			
Vowel perception	81.24 (12.63)	53-75 – 100	.879*
Prosody perception	80.67 (7.58)	56.67 – 90.83	.707*
Linguistic focus	80.63 (10.64)	50 – 100	.605*
Phrase boundary	80.63 (7.98)	57.5 – 100	.276*
Lexical stress	80.75 (11.56)	47.5 – 95	.717*
Speech production	.631(.210)	0 – 1	.254*
Linguistic focus	.613(.195)	.25 – 1	.251*
Phrase boundary	.611(.234)	0 – 1	.209*
Lexical stress	.668 (.197)	.33 – 1	.289*

Uncovering latent variables in behavioural and neural predictors

To minimize the problem of multicollinearity, I used a principal axis factor analysis with varimax rotation (as implemented in fa function from psych R package) to uncover latent variables reflecting shared variance across all behavioural and neural predictors. This was to minimize the multicollinearity problem and reduce the number of predictors entered to the models. The factorability of the dataset was adequate as shown by results of Bartlett's test of sphericity (χ^2=175.85, p<.001), and the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO=.61) was mediocre (Kaiser & Rice, 1974). I extracted 4 factors accounting for 59.46% of variance (17%, 16%, 13% and 13% respectively) following the Kaiser criterion (eigenvalues >1). Factor 1 captured the contributions of dimensional salience in non-verbal stimuli (ITPC non-verbal), Factor 2 – dimension-selective attention to pitch and duration across domains (DSA), Factor 3 – cortical tracking of relative pitch and speech envelope and Factor 4 – dimensional salience in verbal stimuli (Table 16).

Table 16. Factor loadings from exploratory factor analysis

Measure	***Factor 1 ITPC non-verbal***	***Factor 2 DSA***	***Factor 3 mTRF***	***Factor 4 ITPC verbal***
Cumulative variance explained	17%	33%	46%	59%
Dimension-selective attention (DSA)				
Pitch (verbal)	-.08	**.54**	.09	.07
Duration (verbal)	.06	**.71**	.09	.01
Pitch (non-verbal)	.10	**.57**	-.06	.09
Duration (non-verbal)	.19	**.72**	.16	-.13
Dimensional salience (ITPC)				
Pitch (verbal)	.29	.06	.05	**.68**
Duration (verbal)	-.01	-.01	.10	**.84**
Pitch (non-verbal)	**.94**	.04	.10	.08
Duration (non-verbal)	**.50**	.18	.15	.33
Cortical tracking (mTRF)				
Pitch tracking	.15	.14	**1.02**	.08
Envelope tracking	.07	.07	**.71**	.09

Early language experience, attention, salience and cortical speech tracking as predictors of L2 proficiency

The main goal of this chapter was to explore whether and to what extent L2 learning experience, attention, dimensional salience, and cortical speech tracking can explain the differences in L2 language outcomes. I used a series of mixed-effects regression models with L2 learning experience and the extracted four factors as predictors and prosody perception, prosody production, speech perception, vocabulary, and grammar knowledge as outcome measures. First, I entered the L2 language learning experience variables (AOA, LOR and L2 use – Model 1) and then the extracted factors representing dimensional salience in verbal and non-verbal stimuli, attention and cortical speech tracking (Model 2). Model 1 was constructed to examine the extent to which early immersion experience can explain differences in L2 learning outcomes. Model 2 tested the extent to which including the additional variables of attention, salience and cortical speech tracking increases the predictive power of the initial model (i.e., Model 1). The trial-by-trial data from the prosody perception task, speech perception task, vocabulary test and Grammatical Judgement Test were analyzed with separate mixed-effects logistic regressions using the glmer function from the lmer4 package (Bates et al., 2015). The dependent variables were the responses on each trial (0–incorrect, 1–correct). Participants' unique IDs were included as a random intercept along with random slopes for test items. To explore the factors predicting prosody production, I analyzed the production data using the glmmTMB function with a beta distribution appropriate for continuous data bound by the 0-1 interval from the glmmTMB R package (Brooks et al., 2003). The dependent variable was the averaged dichotomized score per participant per linguistic feature. Participants' IDs were included as a random intercept along with random slopes for each feature. The continuous predictors across all models were standardized by centring and dividing by 2 standard deviations using the rescale function from the arm R package (Gelman et al., 2021).

Results of the mixed-effects logistic regression models are presented in Table 17 (see Figure 16 for scatterplots of significant relationships). For prosody perception, attention emerged as a significant predictor (β=.473, p<.001; Figure 16D), indicating that participants who displayed better attention across acoustic dimensions (pitch and duration) achieved more accurate L2 prosody perception. However, this was not the case for prosody production, for which none of the predictors reached significance ($p>.05$ for all predictors in both models). Attention was also found to be a significant predictor of vowel perception ($\beta=.567$, $p=.029$; Figure 16B), along with salience in non-verbal stimuli ($\beta=-.595$, $p=.023$; Figure 16A) and daily L2 use ($\beta=1.106$, $p<.001$; Figure 16C). Adding attention and salience significantly improved the predictive power of the model ($\chi^2=10.385$, $p=.034$). That means that participants' accuracy in detecting vowel differences is tied not only to the intensity of L2 use (the more practice, the better accuracy), but is also positively linked to their ability to attend to acoustic dimensions (the better attention, the better vowel perception) and negatively to their salience (less phase-locking linked to better vowel perception).

I also show that none of the additional variables contributes to explaining L2 vocabulary acquisition. I find that participants' scores on vocabulary tests are solely linked to their LOR ($\beta=-.515$, $p=.007$), however the direction of that relationship seems unexpected as it indicates that better vocabulary knowledge is linked to shorter LOR. However, visual inspection of the scatterplot (Figure 16G) suggests that this relationship is not evident. Finally, I show that for grammar knowledge, daily L2 use emerged as a significant predictor among experience variables ($\beta=.533$, $p=<.001$; Figure 16F) emphasizing the role of daily L2 practice in extracting global prosody information from speech. Furthermore, attention to acoustic dimension was found to contribute to explaining GJT scores ($\beta=.416$, $p=.008$; Figure 16E) and improving the model ($\chi^2=10.482$, $p=.034$) suggesting that participants who use their L2 more on a daily basis and are better able to flexibly attend to acoustic characteristics of auditory inputs also possess better implicit grammar knowledge. Taken together, results from the regression models demonstrated that the significant predictors varied depending on the aspect of L2 language proficiency.

Table 17. Results from mixed-models regression analyses predicting L2 learning outcomes (prosody perception, prosody production, vocabulary knowledge and grammar knowledge). AOA – age of arrival (i.e., the age at which participants arrived in the L2-speaking country for the first time); LOR – length of residence in L2-speaking countries.

	Model 1				**Model 2**			
Prosody perception	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	1.658	.098	16.980	**<.001**	1.657	.091	18.189	**<.001**
AOA	.252	.148	1.704	.088	.220	.125	1.766	.077
LOR	-.245	.148	-1.654	.098	-.184	.133	-1.382	.167
L2 use	.081	.130	.626	.532	.135	.109	1.240	.215
Factor 1 Salience non-verbal	--	--	--	--	-.129	.110	-1.175	.240
Factor 2 Attention	--	--	--	--	.473	.110	4.289	**<.001**
Factor 3 Cortical tracking	--	--	--	--	.145	.108	1.342	.180
Factor 4 Salience verbal	--	--	--	--	.158	.113	1.390	.165
Model comparison	--	--	--	--	χ^2	20.552	**p**	**<.001**
	Model 1				**Model 2**			
Prosody production	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	.613	.094	6.511	**<.001**	.614	.092	6.644	**<.001**
AOA	-.029	.209	-.138	.890	-.025	.208	-.119	.905
LOR	.332	.212	1.562	.118	.391	.224	1.744	.081
L2 use	.271	.186	1.458	.145	.303	.185	1.637	.102
Factor 1 Salience non-verbal	--	--	--	--	-.100	.178	-.562	.574
Factor 2 Attention	--	--	--	--	.217	.183	1.187	.235
Factor 3 Cortical tracking	--	--	--	--	.032	.179	.177	.860
Factor 4 Salience verbal	--	--	--	--	-.084	.188	-.447	.655
Model comparison	--	--	--	--	χ^2	1.955	**p**	.744
	Model 1				**Model 2**			
Vowel perception	**β**	**SE**	z	**p**	**β**	**SE**	**z**	**p**
(Intercept)	2.005	.175	11.482	**<.001**	2.002	.165	12.111	**<.001**
AOA	-.245	.326	-.752	.451	-.256	.300	-.853	.394
LOR	-.189	.319	-.592	.554	-.138	.314	-.438	.661
L2 use	1.025	.289	3.547	**<.001**	1.106	.267	4.144	**<.001**
Factor 1 Salience non-verbal	--	--	--	--	-.595	.263	-2.267	**.023**

Factor 2 Attention	--	--	--	--	.567	.260	2.178	**.029**
Factor 3 Cortical tracking	--	--	--	--	.268	.251	1.068	.286
Factor 4 Salience verbal	--	--	--	--	-.026	.265	-.099	.921
Model comparison	--	--	--	--	χ^2	10.385	**p**	**.034**
	Model 1				**Model 2**			
Vocabulary knowledge	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	1.274	.184	6.935	**<.001**	1.274	.183	6.972	**<.001**
AOA	.359	.192	1.865	.062	.360	.189	1.906	.057
LOR	-.515	.192	-2.691	**.007**	-.519	.201	-2.588	**.009**
L2 use	.227	.167	1.359	.174	.252	.164	1.538	.124
Factor 1 Salience non-verbal	--	--	--	--	-.267	.163	-1.640	.101
Factor 2 Attention	--	--	--	--	.040	.165	.244	.807
Factor 3 Cortical tracking	--	--	--	--	.021	.159	.133	.894
Factor 4 Salience verbal	--	--	--	--	-.108	.169	-.643	.521
Model comparison	--	--	--	--	χ^2	3.266	**p**	.514
	Model 1				**Model 2**			
Grammar knowledge	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	.851	.143	5.956	**<.001**	.851	.139	6.128	**<.001**
AOA	-.190	.194	-.981	.326	-.202	.179	-1.129	.259
LOR	-.136	.195	-.698	.485	-.094	.192	-.487	.626
L2 use	.474	.170	2.788	**.005**	.533	.157	3.397	**<.001**
Factor 1 Salience non-verbal	--	--	--	--	-.305	.156	-1.953	.051
Factor 2 Attention	--	--	--	--	.416	.158	2.632	**.008**
Factor 3 Cortical tracking	--	--	--	--	-.022	.153	-.144	.886
Factor 4 Salience verbal	--	--	--	--	.051	.162	.316	.752
Model comparison	--	--	--	--	χ^2	10.482	**p**	**.034**

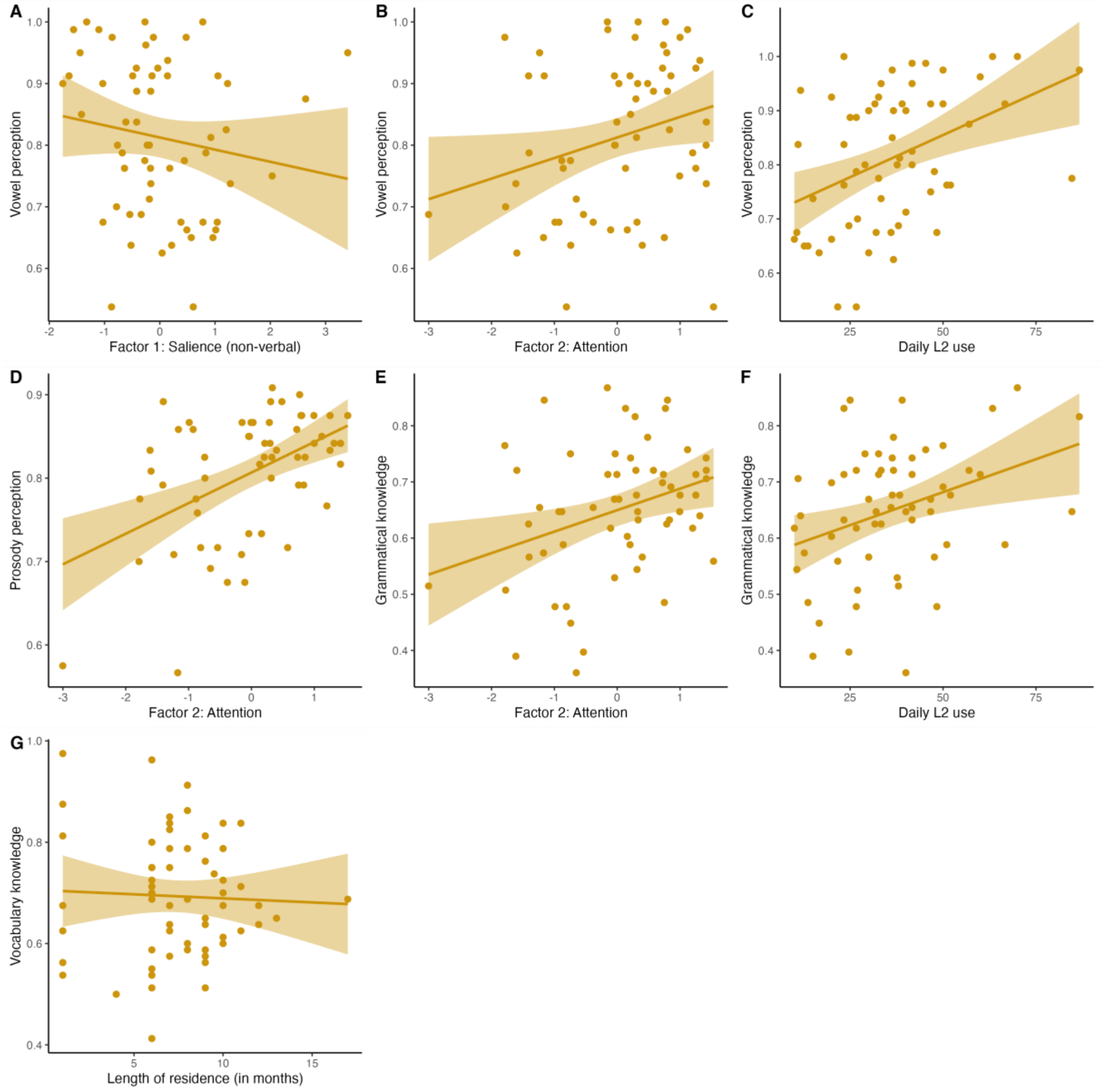


Figure 16. Scatterplots showing significant relationships between L2 proficiency, L2 learning experience, dimension-selective attention, and dimensional salience. (A-C) Vowel perception, (D) Prosody perception, (E-F) Grammatical knowledge, and (G) Vocabulary knowledge. Shaded areas represent 95% confidence intervals.

4.4 Discussion

This study investigated the role of experiential, behavioural and neural factors in predicting L2 learning outcomes across a range of language measures. Based on previous research in second language acquisition (e.g., Yeni-Komshian et al., 1997, 2000), I predicted that better L2 proficiency would be linked to a prolonged and more intensive L2 learning experience. Building on the assumptions of cue weighting theories emphasizing the importance of attention and salience in speech perception (Francis & Nusbaum, 2002), I further predicted that attention and salience would play an additional role in explaining individual differences in L2 spoken language acquisition. These results partially confirmed my predictions regarding the role of the L2 learning experience, showing that more frequent L2 use led to better grammar knowledge and more accurate vowel perception. Supporting my hypothesis about the role of attention in speech perception, attention to acoustic dimensions emerged as a significant predictor of vowel and prosody perception. I also find that attention plays a role in explaining variance in grammatical knowledge. Overall, my findings support the idea that different factors are not only responsible for the development of various aspects of language but also their contributions to boosting L2 learning progress might vary depending on the stage of L2 learning.

Early L2 immersion does not explain individual differences in L2 proficiency

Although ample research exists suggesting that AOA and LOR are significant drivers of L2 proficiency (e.g., LOR was found to have significantly benefited Chinese learners of English perception during the first year of arrival; Li, 2022), in this study, these variables did not predict L2 success. Contradictory to my expectations, I observed a relationship suggesting better vocabulary knowledge with shorter LOR. A potential confound leading to such a result could come from the fact that although participants had relatively short LOR, they all reported many years of EFL education in their home country before arriving in the UK. They could have already acquired a sizeable vocabulary when preparing for their English exams to be able to study abroad. Their knowledge at this level would have been unaffected by this initial LOR or slightly earlier AOA. A more extended immersion would be needed to expand their vocabulary. A perhaps less likely explanation could be that at this level of L2 knowledge, the base vocabulary size might remain stable over a prolonged time. This could be a combined effect of specialized academic training that demands knowledge of vocabulary specific to learners' field of study, refining semantic knowledge in regard to words that have multiple meanings (i.e., expanding the depth of L2 vocabulary knowledge rather than its size), or simply forgetting the words that are not frequently used. It is also worth noting that although the learners achieved, on average, relatively high scores, there was a big discrepancy between how successfully they recognized words and non-words items. Participants could correctly classify most of the real English words, but they still had trouble filtering out similarly looking non-words suggesting the overgeneralization of their word morphology knowledge (McKercher, 2018).

The role of attention in L2 acquisition

Building on the previous research emphasizing the importance of dimension-selective attention and dimensional salience in speech perception (e.g., Francis & Nusbaum, 2002; Holt et al., 2018), I predicted that these variables would emerge as significant predictors of L2 performance across measures related to spoken language. Confirming my expectations, I found that better attention to acoustic dimensions was linked to more accurate vowel and prosody perception. I did not find any link of attention to prosody production – a result that should be interpreted with caution due to low reliability of the employed measure. Nevertheless, this finding aligns with earlier research suggesting that production learning lags behind perception learning (Kuang & Cui, 2018) and could therefore draw on different resources. Other studies also showed that listeners' own speech productions were not predictive of how they weighted acoustic information (relative vs absolute duration, Idemaru, Holt, & Seltman, 2012; VOT vs onset F0, Shultz, Francis, & Llanos, 2012). Despite theoretical differences regarding the basic elements and mechanisms involved in L2 learning, models of L2 speech agree that accurate perception is an essential precursor for accurate production (e.g., Escudero, 2007; Flege, 1995). To examine the nature of the perception-production relationship, further research should carefully consider various aspects of speech perception and production (e.g., examine acoustics of L2 speech production, compare cue weighting patterns during perception vs production; for a detailed review of L2 speech perception-production links and practical guidelines see Schertz & Clare, 2019; Nagle & Baese-Berk, 2022). It is also important to recognize that achieving native-like pronunciation in the target language may not be attainable or necessary for effective communication. While some learners may strive for native-level proficiency, others focus on intelligibility and clear communication rather than eliminating all traces of their foreign accent (Nagle & Huensch, 2020).

My main finding emphasizing the role of attention in vowel and prosody perception is in agreement with the theories of cue weighting (Francis, Baldwin, & Nusbaum, 2000) implicating the role of attention in discerning the categorical membership of newly acquired L2 linguistic categories. The ability to flexibly attend to acoustic dimensions is indeed crucial for putting in focus the relevant acoustic information and weighing its importance to accurately perceive subtle vowel and prosody differences. Since, depending on their language background, listeners might weigh the available acoustic information differently, it is possible that these differences are a contributing factor to difficulties with L2 speech prosody acquisition at segmental and suprasegmental levels – learning an L2 might require people to direct attention to acoustic dimensions that they have grown used to neglecting to learn the distributional statistic of an L2 (Toscano & McMurray, 2010).

The link between attention and L2 acquisition was also found to extend beyond speech perception. Grammatical knowledge was linked to better attention, indicating that the ability to attend to acoustic characteristics of speech might be essential not only for acquisition of L2 speech but also syntax. These results extend the earlier findings showing that grammar was tied to more precise auditory processing (i.e., better temporal processing and synchronization variability and more robust neural encoding of F1, Kachlicka et al., 2019). Both imprecise auditory sensitivity to acoustic details and inability to attend to L2-relevant acoustic dimensions may interfere with the perception of prosodic features since these are conveyed by fine acoustic details. For example, stressed syllables can be realized with a greater pitch movement (e.g., strong vs weak syllable distinction in English, Fear, Cutler & Butterfield, 1995) or be longer (e.g., vowel reduction in Russian, Bondarko, 1977) compared to unstressed syllables. Speakers can employ similar prominence patterns at a word level to mark which part of the sentence or a phrase brings new or important information (Gussenhoven, 2008). Linguistic focus is usually characterized by a greater pitch movement within the emphasized phrase or word (e.g., Breen et al., 2010). Marking where in a long sentence one phrase ends and another begins with appropriate phrase boundaries is achieved by brief changes in pitch and duration (de Pijper & Sanderman, 1994). Difficulties with prosody perception, therefore, could delay the acquisition of syntax, given that listeners can use prosody to segment speech (Shukla, Nespor, & Mehler, 2007; Tremblay, Broersma, & Coughlin, 2018) and detect hierarchical structure in language (Langus, Marchetto, Bion, & Nespor, 2012; Marslen-Wilson et al., 1992). The grammar test included a wide selection of linguistic features, so I could complement my analysis by examining whether any aspects of grammar related to prosody would be more strongly related to attention. It is also possible that experience with L2 speech builds up perceptual expectations about the underlying structure of language that facilitates the processing of the expected grammatical inputs (Tillmann & Poulin-Charronnat, 2010).

Limited role of salience in exploring individual differences in L2 learning

Contrary to my predictions, dimensional salience did not emerge as a significant predictor of L2 success, raising the question of whether cortical tracking of acoustic dimensions can contribute to our understanding of L2 proficiency. As argued in Chapter 2, it is possible that although the ITPC represents tracking of acoustic dimensions within the presented sequences, it does so within sounds that lack resemblance to any natural listening conditions and might not be capturing the acquired salience I intended to measure. However, I measured dimensional salience of acoustic properties during naturalistic speech listening, which was also found to be irrelevant in predicting L2 learning outcomes in my study. Although the analysis technique I used allowed me to describe how stimulus features map onto the neural response, it is possible that the chosen stimulus representations (i.e., speech envelope and pitch) were not appropriate for estimating the relative salience of acoustic dimensions. According to the domain-general auditory encoding hypothesis, neural tracking of speech envelope is driven solely by general auditory mechanisms (Doelling, Arnal, Ghitza, & Poeppel, 2014; Steinschneider, Nourski, & Fishman, 2013). On the other hand, the interactive processing hypothesis argues that cortical tracking of speech envelope reflects the dynamic interactions between the processing of acoustic cues and linguistic information (Zou et al., 2019). In my data, the neural tracking of speech envelope and pitch clustered tightly together during factor analysis, suggesting a common latent factor. However, more evidence is needed to reconcile what the speech envelope tracking represents. If the cortical tracking of speech envelope reflects the mixture of linguistic and acoustic encoding, this measure may not be best suited to reflect dimensional salience. The selection and computation of appropriate stimulus representations for use in encoding and decoding models is a topic of ongoing debate among researchers using naturalistic imaging techniques and, on its own, could be a focus of a fruitful line of methodological research. Further studies should be conducted with the sole aim of understanding various stimulus features and the consequences of their use on the observed neural responses and built models.

Attempts to use cortical tracking of various features as a proxy of L2 proficiency (e.g., word class, speech rate, word position and parts of speech encoding, Ihara et al., 2021; acoustic, phonemic, phonotactic, and semantic features; Di Liberto et al., 2021) show some promise in this respect. For example, di Liberto and collaborators demonstrated clear main effects of proficiency on linguistic encoding (i.e., stronger encoding of L2 phonemic contrasts, stronger and earlier cortical responses to semantic dissimilarity with increasing proficiency; Di Liberto et al., 2021). Perhaps measuring neural entrainment to acoustic (e.g., pitch contour, amplitude envelope) and linguistic features (e.g., syllable and word onsets, phonemic and phonetic representations, semantic similarity) of continuous speech and assessing their individual contributions to explaining L2 proficiency across various metrics could be a feasible way to tackle the questions I asked. For example, better vocabulary knowledge could be linked to better word parsing and encoding of semantic similarity, whereas grammar knowledge could be linked to specific linguistic features exemplifying specific grammatical rules (e.g., plurals, inflections, pauses). Overall, more proficient L2 learners would likely show stronger tracking of linguistic features than acoustic content, more similar to native speakers. It is important to emphasize that when using natural speech, these features might be correlated, so to understand the distinct and shared contributions of these features, I could perform a variance partition analysis (de Heer et al., 2017). This method would allow me to measure the unique variance of each feature (represented by their additional variability brought to the model) and how much they contribute to the model with multiple features.

Conclusions

These findings provide support for cue weighting theories implicating the role of attention in L2 speech acquisition (e.g., Francis & Nusbaum, 2002). I show that attention to acoustic dimensions emerges as a significant predictor of not only vowel and prosody perception, but also grammar knowledge, extending its role from spoken language to syntax. Furthermore, my research contributes to the ongoing discussion about the neurocognitive model of language aptitude (Turker, Seither-Preisler, & Reiterer, 2021; Turker & Reiterer, 2021). These results lend support to the idea that no single factor can explain variance in L2 learning success, as L2 learning is a complex process that encompasses a combination of innate characteristics, experiential factors, cognitive faculties, neural mechanisms and even genetic predispositions (Turker et al., 2021). Integrating attention into this debate is one step towards a more integrative model of L2 proficiency that combines various facets of that process and acknowledges that different factors not only might play a role in mastering different aspects of language but might also be relevant at different stages of L2 learning and have differential effects depending on the individual.

Chapter 5. Effects of targeted perceptual training on cue weighting strategies and L2 speech prosody

Abstract. Speaking a second language (L2) is a highly desirable skill in today's society. However, only some learners achieve the desired level of L2 proficiency, while others struggle to understand L2 speech. Difficulties with learning L2 speech might arise from relying on listening patterns inherited from learners' first language (L1). The importance given to various acoustic dimensions varies across languages, so a strategy optimal for any given L1 might not be useful for subsequent languages. If learning a new language requires redirecting attention to L2-relevant acoustic dimensions, it should be possible to boost L2 learning by enhancing learners' ability to rely on those dimensions. Here, I tested this hypothesis in the context of Mandarin learners of English who tend to overweigh pitch information. I developed perceptual prosody training to help them achieve more native-like strategies by enhancing their ability to use durational cues. 30 learners completed the targeted perceptual training and the control group of 30 participants vocabulary exercises. I measured participants' performance on dimension-selective attention to pitch and duration, dimensional salience of these dimensions and their weighting in prosody categorization tasks before and immediately after the training to test whether their cue weighting strategies changed and if training gains were linked to enhanced attention or salience of these dimensions. Additionally, I measured attention and cue weighting after 6 months to test whether the training had any long-lasting effects on perceptual strategies employed by L2 learners. After the training, participants from the experimental group showed enhanced use of duration specific to categorization of phrase boundaries, where duration is the primary cue. However, these effects did not last over time. Participants in the control group showed more reliance on pitch in categorizing lexical stress. No training-related improvements in prosody perception were observed across tasks. No effects on dimensional salience or dimension-specific effects on attention were found. Overall, these findings offer evidence of strong reliance on pitch among Mandarin learners and difficulties resisting default L1 strategies in L2 speech perception. Nevertheless, I show that adjusting perceptual strategies with targeted training is possible.

5.1 Introduction

Acquiring a new language involves learning new linguistic categories and patterns of how these categories map onto continuous acoustic variations across multiple dimensions. It is a challenging task while learning a first language, but it becomes even more difficult when learning a new language, because acoustic cues can play different roles across languages. For example, native English speakers have not learned to distinguish between various ways of articulating lexical tones marked by changing pitch contours, which, on the other hand, is crucial for conveying meaning in tonal languages (Francis, Ciocca, Ma, & Fenn, 2008; Hao, 2018). In the same way, Japanese native speakers find it difficult to distinguish between the English /r/ and /l/ since there are no such categories in their L1 phonological inventory and they do not show sensitivity to the third formant (F3), which is generally not useful for distinguishing phonemes in Japanese but is crucial for resolving linguistic contrasts in English (Iverson et al., 2003; Ingvalson, Holt, & McClelland, 2012). Furthermore, if L2 learners lack the ability to detect the crucial cue, they might also use cues that are of secondary importance for native speakers (e.g., using duration to distinguish between English vowels /i:/ vs /i/ instead of spectral cues by speakers of various L1s; Polish; Bogacka, 2004; Russian; Kondaurova & Francis, 2008; Mandarin; Flege, Bohn, & Jang, 1997; Zhang, et al., 2015) or give them a greater weight (e.g., relying more on onset F0 in English voicing decisions by Spanish speakers; Llanos, Dmitrieva, Shultz, & Francis, 2013). As a result, an optimal L1 listening strategy might not be as effective for any subsequent L2. Therefore, learning an L2 might lead to changes in the salience of various acoustic dimensions and require people to direct attention to dimensions that can help resolve the new speech contrasts (Francis, Baldwin, & Nusbaum, 2000).

Consistent with the predictions of attention-to-dimension models of perceptual learning, Francis and Nusbaum (2002) showed that improvements in distinguishing between consonant-vowel syllables arise from acquiring similarity within categories and distinctiveness between categories. Both processes are said to be driven by attentional shifts – increased attention to dimensions that best define within-category similarity or differentiate between the categories. Attention might also serve as a perceptual lens, where attending to relevant acoustic dimensions (cue enhancement) and disengaging from task-irrelevant dimensions (cue inhibition) brings that information into focus (Kondaurova & Francis, 2010). However, like the study mentioned above, the vast majority of studies that explored the roles of attention to dimensions and dimensional salience in L2 speech acquisition did so at the segmental level in the context of specific phonetic contrasts (e.g., resynthesized syllables, Baese-Berk, 2019) and minimal pairs (e.g., /sheep/ vs /ship/ and /beat/ vs /bit/ contrasts, Kondaurova & Francis, 2010; /set/ vs /sat/; Liu & Holt, 2015; /Set/ vs /sat/ and /setch/ vs /satch/, Lehet & Holt, 2020; /beash/ vs /peash/, /beak/ vs /peak/, /beef/ vs /peef/, and /beace/ vs /peace/, Zhang, Wu, & Holt, 2021; /beer/ vs /pier/ and /set/ vs /sat/, Wu & Holt, 2022), or even synthesized stimuli that only imitate speech-like acoustic variations (e.g., tones with manipulated carrier and modulation frequency, Roark & Holt, 2019; sequences of nonspeech hums varying in F0 frequency, Obasih, Luhtra, Dick, & Holt, 2023). However, these stimuli do not reflect the full range of acoustic mappings learners need to master as acoustic differences between languages manifest themselves not only at the segmental level, but also at the suprasegmental level (i.e., during prosodic categorization, e.g., phrase boundaries, Kuang, Chan & Rhee, 2022; word stress, Gordon & Roettger, 2017).

Several aspects of prosody are particularly important in learning an L2, namely the syllable prominence within words (lexical stress), emphasizing words containing new or important information (linguistic focus) or placement of phrase boundaries. For example, in English, the position of word stress is not fully predictable like in fixed-stress languages (e.g., primary stress on penultimate syllable in Polish; Wierzchowska, 1971; Domahs et al., 2012; initial syllable stress in Finnish and Hungarian, Peperkamp, Vendelin, & Dupoux, 2010; final syllable stress in standard French, Peperkamp & Dupoux, 2002), and it must be remembered as a part of the pronunciation of all words (Cutler, 2015b). Stressed syllables may be realized in English with a greater pitch movement (strong vs weak syllable distinction, Fear, Cutler & Butterfield, 1995), be longer (Davis & Summers, 1989; Lunden, 2017) compared to unstressed syllables or be marked by an increase in intensity (Plag, Kunter, & Schramm, 2011). Linguistic focus is usually characterized by a greater pitch movement for emphasized phrase or word (e.g., Breen et al., 2010) and phrase boundaries in English are associated with lengthening of the pre-boundary segment (word or syllable) and a rapid change in pitch just before the boundary (Streeter, 1978; Beach, 1991; de Pijper & Sanderman, 1994). But due to inherent differences in cue weighting patterns across languages, English prosody is not realized in the same manner by L2 English learners. For instance, Chinese learners of English were found to rely on pitch cues more than native speakers in categorizing stress patterns in nonsense words, and only pitch was a decisive cue for them (Wang, 2008). In another study, English and Mandarin Chinese native speakers used vowel quality as a primary and pitch as a secondary cue while detecting stress patterns in nonwords, while pitch was completely disregarded by Russian speakers who relied more on duration and intensity cues instead (Chrabaszcz, Winn, Lin, & Idsardi, 2014).

Despite the fact that the suprasegmental features are important for speech comprehension (Cutler, Dahan & van Donselaar, 1997) and were implicated as strong predictors of fluency ratings and degree of perceived foreign accent (e.g., Munro & Derwing, 2001; Trofimovich

& Baker, 2006; Kang, 2010), perceptual strategies underlying acquisition of L2 prosody received relatively little attention in the literature. As a consequence, knowledge about how suprasegmental categories are acquired is rather scarce. In fact, in the context of L2 speech prosody learning, several studies looked at how suprasegmental categories map onto acoustic cues in various languages (e.g., perception of English stress by L1 Mandarin speakers, Zhang & Francis, 2010; perception of English stress by L1 English, Mandarin and Russian speakers, Chrabaszcz et al., 2014; perception of word-final boundaries in French by English and Dutch learners, Tremblay, Broersma, & Coughlin, 2018; perception of intonation in ambiguous English sentences by Korean listeners, Baek, 2022), but there is no research showing how cue weighting in L2 speech prosody changes as a function of L2 learning. To my knowledge, only one study investigated how these mappings can change (Jasmin, Tierney, Obasih, & Holt, 2023). However, these changes were not observed in the context of L2 learning but only due to short-term exposure to artificial accents imposed over emphasized and non-emphasized words, and these effects were only temporary (Jasmin et al., 2023). To test whether relative weights assigned to acoustic dimensions respond to contextual variations, Jasmin and colleagues (2023) asked participants to categorize two-word phrases with emphasis placed either on the first or second word in two contexts –with natural fundamental frequency (F0) and durational variation or with an artificial accent where the acoustic cues covaried atypically. They found that participants adjusted their perceptual weights of primary dimensions to account for perceived irregularities in the input signal but reverted to their default strategies when presented with speech samples with typical accents. These short-term changes help maintain the stability of speech percepts by accommodating temporary fluctuations in cue usefulness (Idemaru & Holt, 2011) or deviations from typical accents (Liu & Holt, 2015; Lehet & Holt, 2017), but are not indicative of long-term learning-induced shifts in attention. L2 learning requires not only adjustments to foreign accents or unknown speakers, but also long-term plasticity that can support re-mapping acoustic dimensions to new linguistic categories. However, it has not yet been tested whether cue weighting patterns can be changed more permanently and lead to improvements in L2 prosody perception.

If indeed L2 category learning relies on adjusting cue weights assigned to acoustic dimensions or is driven by the re-distribution of attention to L2-relevant cues and away from L1-optimal cues, then training the ability to flexibly attend to various acoustic dimensions or increasing their relative salience could be a viable strategy for L2 speech learning. How can L2 learners' attention be redirected from one set of cues to another? The most straightforward way to achieve that goal would be to explicitly instruct participants to pay attention to the specific acoustic dimension relevant to a given language. While such perceptually focused instruction may aid L2 speech learning (e.g., Yang & Sundara, 2019), this kind of instruction is usually not available in naturalistic learning conditions neither while learning a first language nor in most cases of second language acquisition (Francis, Baldwin, & Nusbaum, 2000; Francis & Nusbaum, 2002). It is also possible that due to the differences in cues relevance across languages, learners would not be able to follow such an instruction. This is because they might not show sensitivity to L2-relevant dimensions, or the relative salience of other dimensions is much stronger and attracts their attention instead. For example, Mandarin Chinese speakers were shown to weigh pitch more compared to native English and Spanish speakers and they also struggled to ignore pitch information during speech and non-speech categorization tasks when explicitly asked to do so (phrase boundary and linguistic focus categorization tasks; Jasmin, Sun, & Tierney, 2021). This suggests that since Mandarin speakers are quite adept in using pitch (their default L1 strategy), they might not be inclined to change the strategy that works well for them. Perhaps removing that perceptually overpowering dimension or reducing its reliability is necessary to trigger changes in cue weighting strategies employed by L2 learners. It remains to be seen whether such training-induced changes to perceptual strategies can lead to observable improvements in L2 speech prosody perception.

Present study

I proposed that a major source of difficulty in learning to perceive and produce L2 speech prosody is that individuals have trouble resisting default strategies inherited from their first language. If this is the case, then it could be possible to boost L2 speech learning by training learners to rely on L2-relevant cues that they have grown used to neglecting. First, I asked whether perceptual prosody training can help learners make their perceptual strategies more L2-like. I test this hypothesis in the context of native Mandarin Chinese learners of English who tend to overweigh pitch information in L2 speech perception and production (Zhang, Nissen, & Francis, 2008) and have trouble disengaging attention from pitch even when explicitly asked to do so (Jasmin, Sun, & Tierney, 2021). I designed a training paradigm focused on enhancing the salience and boosting attention to the neglected acoustic dimension – duration. I expected that several days of exposure to acoustically manipulated speech in which pitch was either removed or unreliable, would challenge listeners to use other dimensions than pitch and as a result lead to more flexible perceptual strategies that are not so heavily reliant on pitch. If indeed L2 category learning is accomplished by adjusting cue weights assigned to acoustic dimensions or re-distribution of attention to these cues, then training should increase attention to the cue that listeners are trained to use, while also decreasing attention to the untrained cue (Francis & Nusbaum, 2002). To test whether the targeted training indeed leads to changes in cue weighting and whether these training-induced changes are linked to changes in dimension-selective attention or dimensional salience, I measured participants' performance in attending to pitch versus duration, using these dimensions in L2 prosody categorization tasks and their dimensional salience. I also examined whether training-induced changes in cue weighting and dimensional salience extend to non-verbal perception.

However, it is possible that any changes we observe immediately after the training are only temporary. We know from previous literature that changes to cue weighting strategies can occur as short-term adjustments to atypical accents (Liu & Holt, 2015; Lehet & Holt, 2017; Jasmin et al., 2023), but listers were shown to revert to default strategies when presented with regular speech. It could be that during short-term exposure, the emphasized dimensions are not fully incorporated into learners' listening strategies and so might wane with time. However, the ultimate goal for all L2 learners is to improve their L2 proficiency in the long run, not only temporarily. To test whether the designed perceptual learning intervention has any long-lasting effects on strategies employed by L2 learners, I measured participants' attention performance and cue weighting strategies after 6 months of completing the training. This enabled me to examine whether any initial changes are sustained over an extended period. Understanding the long-term impact of such interventions is vital for discovering the underlying mechanisms and optimizing language learning methodologies.

Finally, I asked whether perceptual strategies training can lead to improvements in L2 prosody perception (i.e., how accurate their categorization of prosodic features is, not what acoustic dimensions they use to do so). The importance of this question is twofold. One is that by examining the influence of perceptual training, I can answer whether it is possible to adjust cue weighting strategies with targeted training. If so, that would open avenues for designing targeted treatments and L2-learning boosting paradigms. However, changing cue weighting strategies seems only relevant if it can lead to long-lasting shifts in attention or ability to flexibly shift attention between acoustic dimensions and if these shifts translate to observable improvements in L2 proficiency. Therefore, the second reason for investigating the impact of perceptual training on cue weighting is to determine whether any changes in cue weighting strategies result in tangible enhancements in learners' prosody perception. By establishing a connection between targeted training and observable progress in L2 learning, I can ascertain the practicality and efficacy of incorporating perceptual training as a valuable component of language acquisition programs.

5.2 Methods

Participants

60 native Mandarin Chinese speakers aged 18-31 (M=22.62, SD=3.27; 53 females, 6 males, 1 non-conforming)[20] took part in the study. They were working or studying in London pursuing degrees in various fields (e.g., psychology, education, economics, law, architecture, or engineering). I recruited the participants from multiple sources, including the SONA platform, social media groups, and communities catering to Chinese students in London (Facebook and WeChat). Mandarin was their primary language, while English was

[20] These are the same Mandarin speakers reported in Chapters 2, 3, and 4.

acquired as a secondary language. The main goal of this study was to offer targeted training to improve participants' listening strategies, so it was crucial to recruit only people with limited immersion experience. They all had learned English as a second language at school and had to pass a recognized English language test (e.g., IELTS, TOEFL, Cambridge English, Person English Test) to enter education in the UK. However, they had little to no experience in daily communication with native speakers of English. Participants were randomly assigned to one of the training groups (Prosody training – experimental, Vocabulary training – control), but their demographics, language and musical background were balanced across groups (Table 18).

Table 18. Summary of basic demographics, musical training and language background information.

	Prosody training group		**Vocabulary training group**	
	M(SD)	**Range**	**M(SD)**	**Range**
Basic demographics:				
Age	22.37 (3.38)	18 – 31	22.87 (3.19)	18 – 30
Gender	25 female, 4 male, and 1 non-conforming		28 female and 2 male	
Music training (> 6 yrs)	N = 13		N = 16	
Immersion experience:				
AOA (age of arrival)[21]	20.8 (3.71)	11 – 29	21.37 (3.37)	14 – 29
LOR (length of residence in months)	7.3 (2.83)	1 – 12	7.52 (3.59)	1 – 17
EFL learning in classroom:				
Age when started	8.13 (3.73)	3 – 18	7.03 (2.09)	3 – 14
Years of training	11.73 (3.35)	3 – 17	13.57 (4.64)	1 – 22
Current L2 use (%):				
In professional settings	65.33 (20.83)	28 – 100	72.83 (25.41)	20 – 100
In social settings	28.43 (22.75)	5 – 95	33.53 (23.87)	4 – 90
At home	3.83 (5.66)	0 – 20	14.23 (24.61)	0 – 80

[21] Both participants who reported their AOA as 11 and 14 only travelled to English-speaking countries at the time (> 1 month) and were not fully immersed in an L2 until they moved to the UK to study.

L2 performance measure

Speech prosody perception

I measured participants' perception of L2 prosody by asking them to categorize spoken phrases representing three prosodic features: linguistic focus, phrase boundary and lexical stress. The stimuli were 120 phrases selected to capture the differences between them (40 phrases; for a full list of stimuli see Supplementary Materials A). All phrases were arranged in pairs representing contrastive instances of a given feature. For example, lexical stress items included two two-syllabic words with stress placed on the first syllable (e.g., PROtest) or the second syllable (e.g., proTEST). To avoid familiarity with the voices they would be listening to during the training, the stimuli for this test were recorded by two different voice actors (male and female; both native British English speakers). All the stimuli were recorded in a sound-proof booth at Birkbeck University with RØDE NT1A large-diaphragm cardioid condenser microphone with shock mount and pop filter and Audacity software (version 3.0.5). The researcher instructed the actors to read the sentences naturally and convey prosodic contrasts as they usually do while speaking.

To make the perception of the prosodic contrast challenging enough to detect individual differences, I morphed two versions of each contrast with the STRAIGHT morphing toolbox for MATLAB (Kawahara & Irino, 2005). I time-aligned these phrases, manually marked corresponding points in both recordings for morphing and created stimuli continua from 0 to 100% with 5% increments across fundamental frequency (F0) and duration. I chose the samples with pitch and duration both set at 70% and 30% for the prosody perception task (values selected based on the pilot study reported in Chapter 4). Selected stimuli were split into two versions of the task in which the presence of male and female voices across features was counterbalanced. Trials were organized in 3 blocks by linguistic features and presented in random order. On each trial, participants heard a spoken phrase or a word and saw two written versions of it on the screen. Their task was to decide whether the spoken phrase or word sounded most like the phrase or word on the left or the right side of the screen. I asked them to respond as quickly as possible without making any mistakes to measure their spontaneous L2 processing (Saito & Plonsky, 2019). The final score was calculated as the proportion of correct responses.

Behavioural measures

Dimension-selective attention task

Base stimuli were eight unique tokens, four for verbal and non-verbal stimuli. To create verbal stimuli, I took two versions of the word "barbecue" from the Multidimensional Battery of Prosody Perception (Jasmin, Dick & Tierney, 2020), one with emphasis placed on that word and one without. I extracted the first vowels /a/ from these words and morphed them along fundamental frequency (F0; other properties were kept constant) using STRAIGHT voice morphing software (Kawahara & Irino, 2005) to create a 100-step continuum. Samples from levels 1 (110.88 Hz) and 56 (124.40 Hz; approximately 2 semitones difference) were selected as stimuli. Then, I manipulated their length (70.58 and 175.83 ms) using Praat software (Boersma & Weenink, 2001). Non-verbal stimuli were 4-harmonics complex tones generated to match acoustic characteristics of verbal stimuli (i.e., 110.91–124 Hz and 70–175 ms). The base stimuli were concatenated into 2 Hz sequences where pitch and duration changed at the same time, but at different rates (0.67 Hz and 1 Hz). Half of the sequences contained repetitions, i.e., instances where a given dimension did not change at the expected rate. There were four types of trials: pitch repetition only, duration repetition only, repetitions in both dimensions and no repetitions in either dimension. The stimuli used in each trial type and attention condition were identical, with the only difference being the focus of attention and were randomly assigned to either attend to duration or attend to pitch conditions and counterbalanced across two versions of the task. Versions were randomized across participants and counterbalanced across testing sessions.

Before each block, participants were told which acoustic property they should attend to. Each trial began with a 500 ms long silence, followed by the sound sequence and a prompt on the screen asking participants whether they detected a repetition within the attended dimension. Participants indicated their responses by clicking the 'Yes' or 'No' button on the screen and were provided with feedback on a trial-by-trial basis. Short breaks were included at the end of each block. The task consisted of 4 blocks for each type of stimulus (verbal and non-verbal), corresponding to two attention conditions and two rates of change. The final score was calculated by combining the hit rates across the dimension change rates for each attended dimension.

Cue weighting tasks

Cue weighting tasks were used to measure participants' reliance on contrasted acoustic cues (pitch vs duration) during categorization of L2 speech prosody (linguistic focus, lexical stress and phrase boundary) and musical beats. In all four categorization tasks, participants were presented with stimuli that varied orthogonally in the extent to which fundamental frequency (F0, correlate of voice pitch) and duration were indicators of one of the two possible linguistic interpretations (for more details about stimuli creation and their acoustic properties see Chapter 2). After listening to each stimulus, participants were asked to categorize the stimuli as belonging to one of two categories: phrase with early or late closure ("If Barbara gives up, the ship" vs "If Barbara gives up the ship"), emphasis on the first or second word ("STUDY music" vs "study MUSIC"), lexical stress on the first vs second syllable ("COM-pound" vs "com-POUND"), and musical beats occurring either every two or three notes ("strong—weak" vs "strong—weak—weak" patterns). There were 10 blocks of each categorization task, which were interleaved in the following order: musical beats, linguistic focus, lexical stress, and phrase boundary.

Pitch and duration weights were derived through the estimation of Firth's biased-reduced logistic regression for individual subjects (Firth, 1993), with pitch and duration levels (ranging from 1 to 4) as predictors of binary responses during categorization tasks. The pitch and duration coefficients were then normalized, resulting in a normalized perceptual weight between 0 and 1, with values closer to 1 indicating stronger reliance on pitch and values nearer to 0 signaling the opposite, and a value of 0.5 indicating an equal reliance on both attributes. The analysis was conducted in R's implementation of Firth's regression in the logistf package (Heinze et al., 2022).

Neural measure

EEG data acquisition

The EEG data were recorded with 32 Ag-Cl active electrodes (standard 10/20 montage), two earlobe reference electrodes and Biosemi™ ActiveTwo system. Data were recorded at a

16,384 Hz sampling rate and digitized with a 24-bit resolution. Impedance was kept below 20 kΩ throughout the testing session. Triggers marking trial onsets (every 6 tones, or 1.2 seconds) were recorded from trigger pulses and sent to the data collection computer. All EEG data processing and analysis were conducted in MATLAB (MathWorks, Inc) using the FieldTrip M/EEG analysis toolbox (Oostenveld et al., 2011) and custom scripts.

Stimuli

The stimuli created for the dimension-selective attention task were base stimuli for the frequency tagging paradigm. They were concatenated into 5-Hz sequences (tone every 200 ms; 96 seconds in total) wherein pitch and duration changed at the same time, but at different rates (every two tones, 2.5 Hz, or every three tones, 1.67 Hz). 20 repetitions (i.e., instances where the dimension did not change at the expected rate) were inserted into each sequence to prevent them from being overly predictable. Within each sequence, 3-5 stimuli were reduced in volume (-12.04 dB) forming amplitude oddballs. Their presence was randomized within each sequence, with the exception that the oddballs could not appear in the initial or final 4.8 seconds (4 epochs) or within 4.8 seconds from another oddball. The same sequences were presented to all participants, but the order was balanced across participants. The stimuli were presented diotically at a maximum of 80 dB SPL, sampled at 44100 Hz using PsychoPy3 (version 3.2.3), and delivered through ER-3A insert earphones (Etymotic Research, Elk Grove Village IL).

Task

Participants listened to sequences and responded to occasional quiet tones via keyboard presses. The main task included 4 blocks with 4 two-minute sequences each. The purpose of this task was to keep participants focused on the sound, but without redirecting their overt attention to pitch or duration. Prior to the main task, participants completed a short practice session outside the EEG booth to make sure they understood the task and could perform well during the main task.

Intertrial phase coherence (ITPC)

The data were resampled to 512 Hz, re-referenced to the mean of the earlobe electrodes, filtered with a sixth-order 20 Hz low-pass and a fourth-order 0.5 Hz high-pass Butterworth filters, and divided into 1.2 seconds epochs. Then, eye blinks and eye movements artefacts were removed based on the independent component analysis (ICA) and visual inspection. A cluster of frontocentral channels (i.e., AF4, F3, Fz, F4, FC1, FC2, FC5, Cz, and C3), with the highest signal amplitudes across all participants were selected for further analysis. The data were averaged across these selected channels. Any remaining artifacts exceeding +/- 100 µV were discarded. To assess cortical tracking of acoustic dimensions, I computed intertrial phase coherence (ITPC) at the frequencies corresponding to the rate of change of acoustic dimensions. Hanning-windowed Fast Fourier Transform was applied to each 1.2-second epoch, then complex vectors at each frequency were converted to unit vectors and averaged across trials. The length of the resulting average vector served as a measure of phase consistency, ranging from 0 (indicating no phase consistency) to 1 (indicating perfect phase consistency). ITPC magnitude at the frequency linked to a specific dimension was used as a measure of salience of a given dimension.

Prosody training (experimental)

The experimental group of Mandarin speakers practiced their ability to perceive prosodic information in English speech. The training stimuli were audio recordings of naturalistic speech capturing various prosodic features – phrase boundary, linguistic focus and word stress. Participants performed categorization exercises[22] where they heard one speech sample at a time, and they were asked to categorise each speech sample as belonging to one of the two categories displayed on the screen.

Stimuli

To create the training materials, I expanded the existing Multidimensional Battery of Prosody Perception (MBOPP, Jasmin et al., 2020) by adding lexical stress stimuli and

[22] Although some suggestions were made in the literature that categorization and discrimination tasks might have different effects (i.e., categorization training might result in acquired similarity that is decrease in sensitivity to within-category differences vs discrimination training may yield an increase in sensitivity to within-category differences; Guenther et al., 1999), research showed that the effects of both types of tasks are comparable (e.g., Flege, 1995; Wayland & Li, 2008).

additional recordings from multiple speakers for linguistic focus and phrase boundary stimuli. The final dataset consists of recordings made by six professional voice actors (3 males, 3 females) to guarantee a high variability of speakers' speech rate (female voices were 116.24, 108.4 and 127.19 words per minute, M_{female}=117.39; male voices were 175.62, 136.02, and 140.06 words per minute, M_{male}=150.57), pitch range (females voices were 229.51, 248.77 and 206.65 Hz, M_{female}=228.31 Hz; male voices were 145.18, 111.27 and 123.83 Hz, M_{male}=126.76 Hz), and age (females were 24, 33 and 43 years old, M_{female}=33.33; males were 25, 36 and 46 years old, M_{male}=35.67) during training.

Stimuli recording

First, all the actors were asked to submit one sentence recording to evaluate the quality of their audio setup. After receiving a short audio snippet from each actor, the researcher listened to this sample and examined the spectrograms to check the audio recording quality (e.g., the degree of background noise, volume, closeness to the microphone). This step aimed to ensure that the recording conditions and the audio quality of new recordings would be comparable with the original recordings from the MBOPP dataset. All samples were recorded as high-resolution two-channel wav files with a sampling rate of 44.1 kHz and 16-bit audio bit depth.

After completing this step, actors attended a training session via Zoom during which the researcher explained the purpose of the project and gave instructions as to how they should record the stimuli. After a short rehearsal with the researcher, actors were provided with a complete list of stimuli to read and record. This list included up to 100 sentences for each of the three prosody tests – 84 sentences for "Phrase Boundary", 94 sentences for "Linguistic Focus", and 100 sentences for "Lexical Stress" (for the complete list of stimuli see Supplementary Materials C). The sentences were arranged into pairs forming target prosodic features contrasts. For example, in lexical stress stimuli, one sentence had a word with stress placed on the first syllable (e.g., "Most buildings in the COM-pound are connected by tunnels"), and the second sentence in the pair included a word with stress placed on the second syllable (e.g., "Changes to current policies will only com-POUND the problem"). All speakers were asked to read the sentences aloud using their usual strategies to convey the above-mentioned prosodic features. The speakers were asked to make sure that their speech sounds as natural as possible while emphasising the target contrast (for the detailed instructions given to voice actors see Supplementary Material D).

Stimuli processing

All wav files were first trimmed to remove any silences at the beginning and end of each recording and downmixed to mono by averaging the existing channels. An additional step involved extracting target words from the carrier sentences for stress stimuli, as well as extracting identical portions of recordings for linguistic focus and phrase boundary stimuli. Then, all samples were normalised to a target loudness level of -20 dB RMS[23] with the "Match loudness" function in Adobe Audition software (Adobe Inc.). Voice stimuli used in the training were then processed using STRAIGHT voice morphing software (Kawahara & Irino, 2005) by morphing two target contrast recordings along two acoustic dimensions – fundamental frequency (F0, correlate of voice pitch) and duration. For all samples, pitch was set at 50% (i.e., representing mid-values of F0 between the two contrast recordings), and duration was increased from 0% to 100% in 5% increments. For example, the resulting morphs of words "COM-pound" (stress placed on the first syllable) and "com-POUND" (stress placed on the second syllable) correspond to 0% duration represent one end of the continuum (i.e., the word "com-POUND", with second stressed syllable longer than the first) and stimuli with 100% duration represent the other extreme of the continuum (i.e., the word "COM-pound", with first stressed syllable longer than the second). The stimuli with all the remaining values of duration are perceptually the in-between versions that change smoothly from "com-POUND" to "COM-pound" and duration cue was meant to be the only cue that can reliably suggest these categories. In English, pitch and duration covary orthogonally such that longer duration and higher pitch co-occur and signal placed emphasis or lexical stress whereas shorter duration and lower F0 values are associated with lack of emphasis or stress.

[23] The total RMS values in -dB calculated by Audition to normalize speech or music stimuli are relative to the average of all waveforms, not absolute. They were used only to equalize the volume across samples coming from different speakers, not to control their output volume.

Whispered speech was created by transforming original recordings with the "Whisper" command from the Praat Vocal Toolkit (Corretge, 2023). After transformation, there were no sufficient traces of voicing in the recordings to detect F0.

Acoustic similarity analysis

Acoustic similarity analysis of experimental stimuli was conducted to flag target contrasts with high estimated similarity that could be potentially difficult for listeners to distinguish from one another. The analysis was performed on all target contrasts by comparing the amount of dynamic time warping (DTW) needed to morph the samples onto one another. The DTW computation was done using custom MATLAB scripts and continuous dynamic time warping function (Mico, 2023). The obtained value of 0 means that no DTW is required to map two speech samples onto one another (i.e., the samples are identical). The bigger the value, the more significant the difference between the signals and their expected distinguishability. Based on pilot listening to all stimuli with durations values of 0% and 100 % that at least theoretically should be easy to distinguish, I established that the differences between stimuli with distances lower than 5 might be difficult to perceive. Therefore, all the sound samples with distance values lower than 5 were marked for manual inspection. The native English researcher listened to all the flagged recordings and decided whether the difference between them was audible enough to detect. Inaudible contrasts were excluded from the training materials. The final stimuli set consisted of 5416 focus stimuli, 5572 phrase stimuli and 5076 stress stimuli that were allocated to appropriate difficulty levels depending on their durational cue contents.

Adaptive difficulty levels

The main goal of the training was to help participants achieve more English native-like perceptual strategies by enhancing their ability to use various acoustic cues during listening. Since Mandarin speakers tend to overweight pitch, the training emphasised duration as the most reliable cue and downplayed the role of pitch cues to balance their importance. I gradually introduced task-irrelevant pitch variation and decreased the size of the duration cue to increase the difficulty of the tasks throughout the training. By doing so, I wanted to balance participants' use of pitch and duration to a more native-like level and increase their flexibility in using acoustic information in speech perception.

Participants start the training by listening to whispered speech, which does not contain any voiced elements. Removing the otherwise salient pitch from the initial levels of training facilitates listening to the remaining acoustic information. It also allows the listeners to focus on duration when making categorical judgements, as no competing information is available. Then, they move on to listening to transformed voiced speech. In these samples, pitch contour is audible and changes naturally within speech, but it is informatively ambiguous (i.e., it does not help the listeners resolve the target contrast as its values were set at 50% between the two contrasting recordings). It is to keep the pitch information present and force the listeners to rely on other information when making categorical decisions. An additional layer of difficulty is introduced by decreasing the step size of the duration information. The easiest levels contain the target contrasts with the extreme duration range (e.g., 0% vs 100%, 5% vs 95%, which are much more easily distinguishable), and more challenging levels that include samples getting closer to each other in the stimuli space (e.g., 20% vs 80%, 30% vs 70%) are introduced as participants proceed through the training. The closer to the centre of the stimulus space the samples are, the more ambiguous they become. Table 19 provides an overview of the main difficulty levels and stimulus properties included at each level.

Table 19. General description of training setup and distribution of acoustic information (pitch and duration) across difficulty levels for each linguistic feature. Samples with durations sampled from the middle of the stimuli space (i.e., 50% vs 50%) and those close to the middle of the distribution (i.e., 45% vs 55% and 40% vs 60%) were not included due to their high perceptual ambiguity. Each level was as a separate set containing all the possible trials for a given level – from 40 to 150 speech samples from which the trials are drawn. There were 30 levels in total for each linguistic feature.

	Levels 1 – 3	**Levels 4 – 30**
Pitch	No pitch	Ambiguous pitch
Importance	No pitch information interfering with informativeness of durational cues	Pitch kept at neutral middle-point providing no information about the categorical membership
Information	0 %	50 %
Duration	Natural duration	Multiple levels of duration from long to short (increasing difficulty)
Importance	Duration as the only reliable cue	Duration as the only reliable cue, but the size of the duration cue decreases with increasing levels
Information	100 %	0 vs 100 % (levels 4 – 6) 5 vs 95 % (levels 7 – 9) 10 vs 90 % (levels 10 – 12) 15 vs 85 % (levels 13 – 15) 20 vs 80 % (levels 16 – 18) 25 vs 75 % (levels 19 – 21) 30 vs 70 % (levels 22 – 24) 35 vs 65 % (levels 25 – 27) 40 vs 60 % (levels 28 – 30)

Vocabulary training (control)

The control group of Mandarin speakers practised their English skills with vocabulary boosting exercises matched in length and intensity with the experimental training. Vocabulary training was chosen to satisfy the requirements of an effective control condition. Namely, it was essential to select an active control condition that would not overtly appear as such (e.g., no practice or completing tasks not related to L2 learning would immediately be recognized as control condition) and that both training paradigms would be equally engaging and motivating for participants as such differences could have significant consequences for completing the project (e.g., participants in no training group could have been less motivated to complete the study) or interpreting the results (e.g., observed effects could be driven by lack of motivation to complete the tasks by participants assigned to control group). In the vocabulary training, participants could practice their knowledge of English words and learn increasingly more challenging vocabulary. Participants were asked to match pictures, words, and short phrases with their translations. They saw an image, a word in Chinese or English, and they had to click on the appropriate English word to name the object, translate the Chinese word or match the English word with its short definition.

Stimuli and difficulty levels

The main vocabulary selection used in the control training was taken from the Vocabulary Test developed by Schmitt (2012). These materials include lists of words arranged based on their frequency of occurrence and reflect the order in which the L2 learners would most likely acquire the vocabulary (i.e., vocabulary profile; Nation, 2006; Cobb, 2012). For example, the words from 2K level would be learned first by English learners since these words are most frequently used. Then, they would learn words from the subsequent groups. The training stimuli included words from the following vocabulary levels: 2K level words (e.g., birth, dust, operation), 3K level words (e.g., assist, bother, condemn), academic vocabulary (3K and 4K level words, e.g., anticipate, principle, empirical), 5K level words (e.g., casual, desolate, fragrant) and 10K level words (e.g., smoulder, luscious, primeval). For the full list of stimuli see Supplementary Materials E.

Within each of word level, participants completed 6 different types of tasks with different levels of complexity (Table 20). Introducing various tasks served two functions – one, to make the training more engaging and two, to make the tasks more challenging without adding new vocabulary. Using different tasks with the same vocabulary facilitated the repetition of new material and allowed participants to test their knowledge in different contexts (i.e., naming, translation, and definition). First, participants completed picture-matching tasks. They saw one picture and needed to select which word best described that picture, or they saw three images and they had to decide which picture corresponded to the word displayed on the screen. These tasks included only concrete words (i.e., primarily nouns or adjectives) that could be easily depicted with an image. Apart from target words, each trial included two filler words. The words used as fillers were selected words with related meanings (not synonyms), minimal pairs, words that sound similar (homophones) to target words or closely matched rhymes (for examples see Table 21). For example, if I had a target word “shoe”, good filler words could be “sue” (similarly sounding) and “sock” (related meaning – you put both on your feet). But “boot” or “sneaker” would not be good examples as these are synonyms of “shoe”. Next, participants moved on to word-to-word translations, where they had to either match one of the English words with one of the three Mandarin Chinese translations or select an appropriate English translation from the three options given for a single Chinese word. In these tasks, I introduced more abstract words and verbs that could not be easily portrayed with an image. I used the same filler words as in the picture task. The last tasks included trials in which participants needed to match English words with their short definitions written in English.

Table 20. Description and examples of vocabulary training tasks.

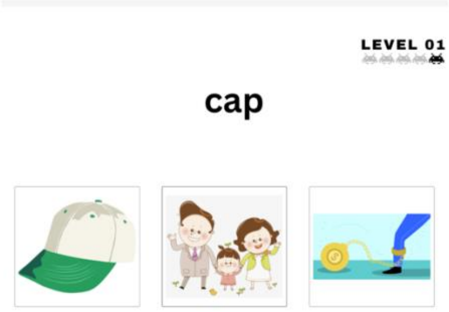
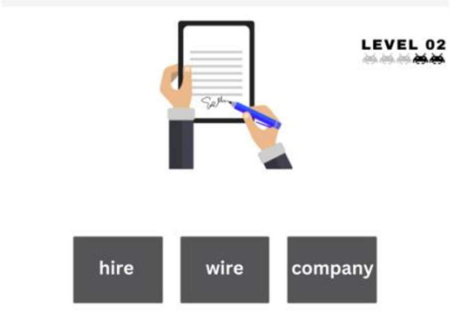
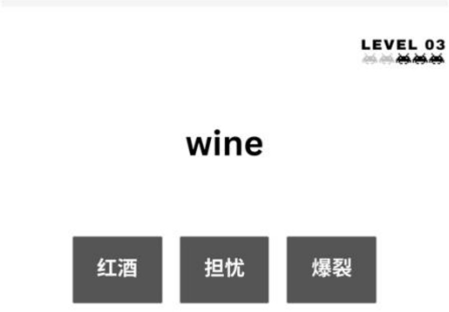
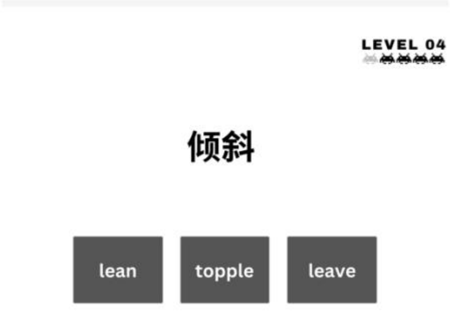
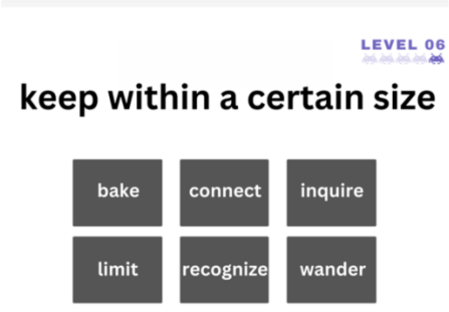
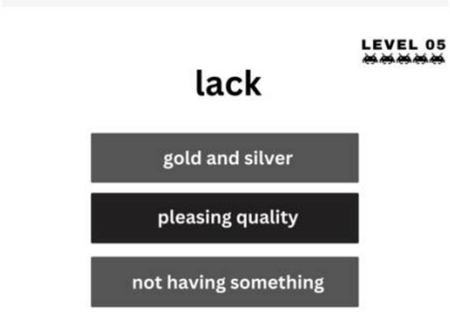
Type of task	**Task 1**	**Task 2**
Pictures	Participants see three images and one English word and need to decide which picture best represents the word	Participants see one picture and three English words and need to select which word corresponds to the picture
Examples		
Word-to-word translations	Participants see three words in Chinese and one word in English and need to select the correct translation	Participants see one word in Chinese and three words in English and need to select the correct translation
Examples		
Definitions	Participants see a list of English words and need to select the word that best fits the definition in English	Participants see three definitions in English and need to choose which one describes the English word
Examples	 LEVEL 06 keep within a certain size bake connect inquire limit recognize wander	 LEVEL 05 lack gold and silver pleasing quality not having something

Table 21. Examples of filler words used for trials with pictures and word-to-word translations.

Target word	Semantic filler (meaning-related)	Phonological filler (pronunciation-related)
media	microwave	median
vehicle	mechanic	pickle
adult	child	assault
topic	converse	tonic

All the words included in the training were translated to Mandarin Chinese by a linguistically trained native speaker of Mandarin Chinese with several years of professional experience in conducting English to Chinese and Chinese to English translations. Where the translator encountered English words that had two or more different meanings in English, each of which was linked to a different word in Mandarin Chinese (for example, "adopt" can be translated as "adopt a child" or "adopt an approach"), the translator consulted the "commonality of use" (i.e., which meaning is more frequently used in Mandarin Chinese) and used the more commonly used meaning for translation. Supplementary Materials E include all target vocabulary and filler items used in the training materials.

Training procedure

Both training paradigms were designed and hosted on the Gorilla platform (Anwyl-Irvine et al., 2020). The design included features drawn from video games (e.g., levels display, color themes, audio feedback) to maximize participants' engagement. Participants completed the same tasks every day for 6 days (approximately 30 minutes a day). Although I recommended practicing every day, I allowed for up to 2 days of breaks during the training (i.e., participants completed 6 training sessions across 8 total days). Compliance was monitored daily by the researcher. Each session included 15 blocks of 20 trials of exercises (300 trials per day, 1800 trials in total). In prosody training, there were 5 blocks for each linguistic feature. Scores and level advancements were computed separately for each linguistic feature (linguistic focus, phrase boundary and lexical stress) so that between-feature differences did not interfere with participants' overall progress. This is because duration might be more or less informative across features, which could prevent participants from progressing to the next levels if they had failed one linguistic feature.

Participants could always see which feature they were practising in the top left corner of the screen. The order in which the feature blocks were presented was randomised across training sessions and between participants. In vocabulary training, there were 15 blocks with 20 trials of vocabulary exercises per day. For both types of training, each set of 20 trials presented in each block was randomly selected for each participant from that stimulus set. If a participant did not reach the pass threshold and had to repeat a given level, the algorithm selected another random selection of 20 trials from that level (i.e., participants did not see the same trials in the same order again).

Both forms of training were adaptive, becoming gradually more difficult when participants performed well to ensure that the tasks remained challenging. An adaptive training regimen was meant to provide a good balance between easiness and difficulty. If the tasks were too difficult (e.g., the duration difference between the stimuli would be too small to perceive), then the training would not be effective because participants would not be able to detect the cue they should rely on to complete the task. Similarly, introducing advanced vocabulary to beginner learners would require them to learn a lot of new words too quickly and perhaps be very frustrating. On the other hand, if the tasks were too easy, they would not present any challenge to participants and would not encourage learning. This, in turn, could result in a lack of motivation and higher dropout rates. Adaptiveness allows participants to quickly move through the easier levels and spend time on challenging tasks where more practice is needed to elicit change. The levels participants achieved at the end of the training might vary, but the most relevant is that they completed the same amount of training and were consistently challenged. Especially in the Prosody Group this balance is crucial, as the perceived salience of acoustic dimensions might vary across participants. With an adaptive procedure, the stimuli they heard and the levels they reached represent the limits of how difficult the task could be for them.

After each block, participants received feedback on how well they performed during that block. If they scored 75% correct or more, they moved up a level. They had to repeat the current level if they scored below 75% (i.e., made more than 5 mistakes). That means that although participants completed the same amount of training (i.e., the same number of trials), the number of levels they cleared depended on their performance since the training was adaptive. Their progress carried over across training days. Additionally, immediate feedback was provided on the screen for each trial, and a summary of scores was given at the end of each block. Participants could track their progress within each block by looking at the progress bar displayed on the top of the screen. They could see their current level throughout the training displayed in the top right corner of the screen. Participants also obtained a training brochure explaining the training set up and procedure (see Supplementary Materials F). Processed data, analysis scripts and Supplementary Materials can be found at: *https://osf.io/5j8pe/?view_only=874b45bef48c4839807867f12a02809a*. Online materials used to run both types of training are available for preview at: *https://app.gorilla.sc/openmaterials/580460.*

General procedure

Interested participants who responded to study adverts were invited to a short telephone or video call. The purpose of this call was to check whether they met all the study criteria and answer participants' questions about the study procedures and its purpose. After giving their informed consent, participants began the study. The study consisted of four parts. Part 1 (Pre-Test) consisted of a series of online tasks (detailed demographics questionnaire, dimension-selective attention task, categorization tasks and speech prosody perception task) and EEG session in the lab at the Department of Psychological Sciences at Birkbeck, University of London. Part 2 (Training) involved 6 days of online language training (experimental prosody or control vocabulary). After the training, Part 3 (Post-Test I) included the same online tasks and a second EEG session in the lab. Finally, Part 4 (Post-Test II) took place 6 months after the end of Part 3 and involved completing the same online tasks once again. Participants were not invited for another EEG session after 6 months because I did not find any significant effects in the EEG data at Post-Test I and, therefore, did not expect to observe any changes at Post-Test II. From 60 participants who completed Parts 1, 2, and 3 who were contacted after 6 months, 42 completed Part 4. The data presented in this chapter includes the behavioural and neural data from Mandarin speakers at Time I, Time II and Time III (I.e., at Pre-Test, Post-Test I and Post-Test II; see Figure 1). All the ethics procedures were approved by the departmental Ethics Committee.

All participants were reimbursed for their time in cash (at £10 per hour) or its equivalent in course credits.

5.3 Results

To test whether Mandarin speakers can be trained to make use of other acoustic dimensions than pitch and make their strategies more native-like, I tested their performance on a dimension-selective attention task, cue weighting during speech and musical beats categorization tasks and assessed dimensional salience of pitch and duration before and after the training. To assess the immediate training effects, I built a series of regression models, with training group (Prosody, Vocabulary) and testing time (Pre-Test, Post-Test I) as predictors. To test whether training has any long-term effects, I measured dimension-selective attention and cue weighting strategies after 6 months, and I built models with training group (Prosody, Vocabulary) and an additional testing session (Pre-Test, Post-Test I, Post-Test II) as predictors.

Effects of training on dimension-selective attention

The data from the dimension-selective attention task were analyzed with a mixed-effects regression model. I used the glmmTMB function from the glmmTMB package (Brooks et al., 2003) with a beta distribution appropriate for the format of attention data – the proportions of correct responses take values from 0 to 1. I built two models. The first model tested the short-term training effects and included data from Pre-Test and Post-Test I that took place immediately after the training (Model 1, N=60). The second model examined whether training effects have any long-term effects and included data from Pre-Test, Post-Test I and additional Post-Test II after 6 months (Model 2, N=42). The categorical variables representing training group (vocabulary, prosody), domain (speech, tones) and attended dimension (duration, pitch) were treatment coded, with vocabulary training group, speech stimuli and duration serving as a baseline and prosody training group, tone stimuli and pitch as a group comparison (0 and 1, respectively). In Model 1, the testing session (Pre-Test, Post-Test I) was treatment coded (0 and 1, respectively). In Model 2, I used the contr.sdif function from the MASS package (Ripley et al., 2023) to estimate successive differences between Pre-Test (Time I) and Post-Test I (Time II) and then between Post-Test I (Time II) and Post-Test II after 6 months (Time III). Participants' unique IDs were included as a random intercept. Including random slopes for domain and dimension resulted in overfitting, so simpler models were used. The main effects of interest were the three-way interactions between the training group, testing time and dimension or domain. Results of both models are presented in Table 22 and Figure 17.

A three-way interaction between time, group, and domain from Model 1 suggests that the relationship between time and group differs between speech and tone stimuli (β=.98, p=.024). A four-way interaction between time, group, domain, and dimension also emerged as significant (β=-1.33, p=.046). To follow up on these interactions, I built four separate regression models predicting attention performance with time and group as predictors for all four task conditions (speech pitch, speech duration, tones pitch and tones duration). None of the post hoc tests revealed significant effects (p>.05). No significant changes in dimension-selective attention were observed in Model 2 (p>.05), suggesting no long-term training improvements.

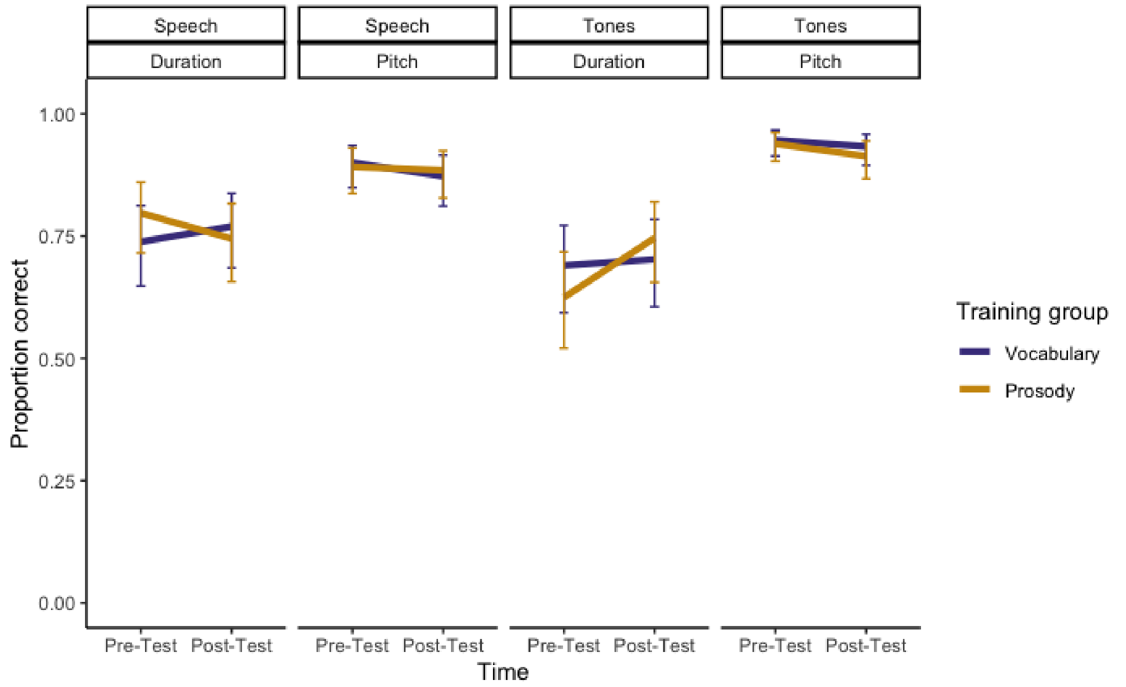


Figure 17. Predicted proportion of correct responses in dimension-selective attention task at Pre-Test and Post-Test I for Prosody and Vocabulary training groups. Responses averaged across participants; error bars – 95%CI.

Table 22. Results from mixed-effects regression analyses testing the differences between Pre-Test and Post-Test I (Model 1) and Pre-Test, Post-Test I and Post-Test II (Model 2) in dimension-selective attention.

	Model 1				**Model 2**			
Predictor	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	1.038	.219	7.748	**<.001**	.700	.201	3.481	**<.001**
Time (I → II)	.169	.216	.782	.434	.201	.255	.787	.431
Time (II → III)	--	--	--	--	-.297	.253	-1.170	.242
Group (Vocabulary)	.331	.314	1.051	.293	.456	.287	1.591	.112
Domain (Speech)	-.239	.210	-1.139	.255	-.159	.145	-1.097	.272
Dimension (Duration)	1.162	.236	4.919	**<.001**	1.087	.160	6.803	**<.001**
Time (I → II) x Group	-.467	.308	-1.517	.129	-.298	.372	-.800	.423
Time (II → III) x Group	--	--	--	--	-.070	.366	-.193	.847
Time (I → II) x Domain	-.108	.302	-.358	.720	-.030	.352	-.085	.932
Time (II → III) x Domain	--	--	--	--	.664	.355	1.873	.061
Group x Domain	-.621	.306	-2.030	**.042**	-.175	.210	-.835	.403
Time (I → II) x Dimension	-.448	.328	-1.368	.171	-.646	.387	-1.670	.095
Time (II → III) x Dimension	--	--	--	--	.482	.383	1.258	.208
Group x Dimension	-.421	.337	-1.250	.211	-.158	.227	-.694	.488
Domain x Dimension	.902	.335	2.694	**.007**	.880	.267	3.881	**<.001**
Time (I → II) x Group x Domain	.977	.431	2.265	**.024**	.807	.513	1.573	.116
Time (II → III) x Group x Domain	--	--	--	--	-.491	.513	-.958	.338
Time (I → II) x Group x Dimension	.673	.466	1.446	.148	.937	.559	1.677	.094
Time (II → III) x Group x Dimension	--	--	--	--	-.399	.552	-.722	.470
Time (I → II) x Domain x Dimension	.160	.472	.340	.734	.304	.559	.547	.585
Time (II → III) x Domain x Dimension	--	--	--	--	-1.051	.551	-1.907	.056
Group x Domain x Dimension	.577	.476	1.212	.225	-.249	.323	-.770	.441
Time (I → II) x Group x Domain x Dim	-1.329	.667	-1.994	**.046**	-1.523	.794	-1.918	.055
Time (II → III) x Group x Domain x Dim	--	--	--	--	.828	.786	1.053	.292

Effects of training on dimensional salience

To explore whether targeted training leads to any changes in neural tracking of acoustic dimensions, I analyzed the ITPC data with the mixed-effects regression model. I used the glmmTMB function from the glmmTMB R package (Brooks et al., 2003) with a beta distribution appropriate for continuous ITPC data bound by the 0-1 interval. The dependent variable was the mean ITPC. The categorical variables representing the testing session time (Pre-Test, Post-Test I), training group (vocabulary, prosody), domain (speech, tones), and acoustic dimension (duration, pitch) were treatment coded (0 for reference level, 1 for comparison level). Participants' unique IDs were included as a random intercept. The main effects of interest were the three-way interactions between the training group, testing time and dimension or domain. Results from the mixed-effects regression model (Table 23) demonstrated no training improvements (p>.05).

Table 23. Summary of effects in mixed-effects regression model testing dimensional salience at Pre-Test and Post-Test I.

Predictor	***Estimate***	***SE***	***z***	***p***
Intercept	-1.635	.067	-24.255	**<.001**
Time (Pre-Test)	-.117	.074	-1.598	.110
Group (Training)	-.054	.096	-.566	.571
Domain (Speech)	-.390	.077	-5.043	**<.001**
Dimension (Duration)	-.534	.080	-6.679	**<.001**
Time x Group	.083	.104	.797	.425
Time x Domain	.128	.110	1.157	.247
Group c x Domain	-.036	.111	-.320	.749
Time x Dimension	-.107	.117	-.920	.358
Group x Dimension	-.093	.115	-.840	.422
Domain x Dimension	.450	.116	3.888	**<.001**
Time x Group x Domain	-.083	.158	-.529	.597
Time x Group x Dimension	-.064	.168	-.383	.702
Time x Domain x Dimension	.078	.166	.467	.641
Group x Domain x Dimension	.132	.166	.792	.428
Time x Group x Domain x Dimension	.099	.239	.415	.678

Effects of training on cue weighting strategies

Immediate training effects

To quantify whether listeners' use of acoustic cues in categorization changed as a result of training, the trial-by-trial categorization data were analyzed with a series of mixed-effects logistic regression models using the glmer function from the lme4 package (Bates et al., 2015). The dependent variable was the response on each trial (0–incorrect, 1–correct). The categorical variables representing the testing session time (Pre-Test, Post-Test I) and training group (Vocabulary, Prosody) were treatment coded with the first variable level serving as a baseline and the second as a group comparison (0 and 1 respectively). The continuous predictors pitch level (1-4) and duration level (1-4) were standardized by centering and dividing by 2 standard deviations using the rescale function from the arm R package (Gelman et al., 2021). The resulting beta coefficients from the model represent the change in log odds given an increase of one standard deviation of that variable. Participants' unique IDs were included as a random intercept. Including random slopes for pitch level and duration level and their interaction resulted in overfitting, so the simpler models without random slopes were selected across categorization tasks. The main effects of interest were the three-way interactions between the training group, testing time and use of pitch or duration. Results of the mixed-effects logistic regression models are presented in Table 24.

Table 24. Summary of effects in mixed-effects logistic regression models for language and musical beats categorization tasks at Pre-Test and Post-Test I.

Task	***Predictor***	***Estimate***	***SE***	***z***	***p***
Linguistic Focus					
	Intercept	-.040	.142	-.278	.781
	Time (Pre-Test)	.248	.067	3.714	**<.001**
	Group (Vocabulary)	.378	.202	1.871	.061
	Pitch	5.350	.140	38.120	**<.001**
	Duration	.339	.090	3.744	**<.001**
	Time x Group	-.089	.091	-.923	.356
	Time x Pitch	.821	.211	3.899	**<.001**
	Group x Pitch	.437	.207	2.109	**.035**
	Time x Duration	.107	.133	.802	.422
	Group x Duration	.269	.132	2.040	.041
	Pitch x Duration	.344	.271	1.269	.204

Time x Group x Pitch	-.426	.300	-1.417	.157
Time x Group x Duration	-.290	.192	-1.509	.131
Time x Pitch x Duration	-.945	.418	-2.260	**.024**
Group x Pitch x Duration	.318	.396	.804	.421
Time x Group x Pitch x Duration	.554	.596	.930	.353
Phrase Boundary				
Intercept	-.331	.076	-4.376	**<.001**
Time (Pre-Test)	.089	.056	1.593	.111
Group (Vocabulary)	-.091	.107	-.851	.395
Pitch	1.654	.080	20.678	**<.001**
Duration	2.897	.088	22.032	**<.001**
Time x Group	.223	.078	2.872	**<.004**
Time x Pitch	.261	.114	2.284	**<.022**
Group x Pitch	.226	.112	2.018	**<.044**
Time x Duration	-.060	.123	-.489	.625
Group x Duration	-.651	.119	-5.487	**<.001**
Pitch x Duration	-.139	.172	-.810	.418
Time x Group x Pitch	-.247	.159	-1.554	.120
Time x Group x Duration	.424	.169	2.514	**.012**
Time x Pitch x Duration	-.140	.242	-.578	.564
Group x Pitch x Duration	.214	.235	.910	.363
Time x Group x Pitch x Duration	-.078	.334	-.234	.815
Lexical Stress				
Intercept	-.306	.129	-2.376	**.018**
Time (Pre-Test)	-.176	.071	-2.490	**.013**
Group (Vocabulary)	.098	.182	.539	.590
Pitch	5.692	.151	37.581	**<.001**
Duration	.493	.095	5.201	**<.001**
Time x Group	.268	.097	2.755	**.006**
Time x Pitch	.982	.232	4.237	**<.001**
Group x Pitch	.200	.217	.923	.356
Time x Duration	.020	.141	.144	.886
Group x Duration	.038	.135	.279	.780
Pitch x Duration	.467	.293	1.597	.110
Time x Group x Pitch	-1.289	.310	-4.153	**<.001**
Time x Group x Duration	-.061	.194	-.315	.753
Time x Pitch x Duration	.123	.460	.266	.790
Group x Pitch x Duration	-.032	.418	-.078	.938
Time x Group x Pitch x Duration	-.261	.617	-.423	.672
Musical Beats				
Intercept	-.374	.212	-1.763	.078
Time (Pre-Test)	.091	.085	-1.073	.283
Group (Vocabulary)	.281	.302	.930	.352
Pitch	8.697	.257	33.798	**<.001**
Duration	1.921	.129	14.940	**<.001**

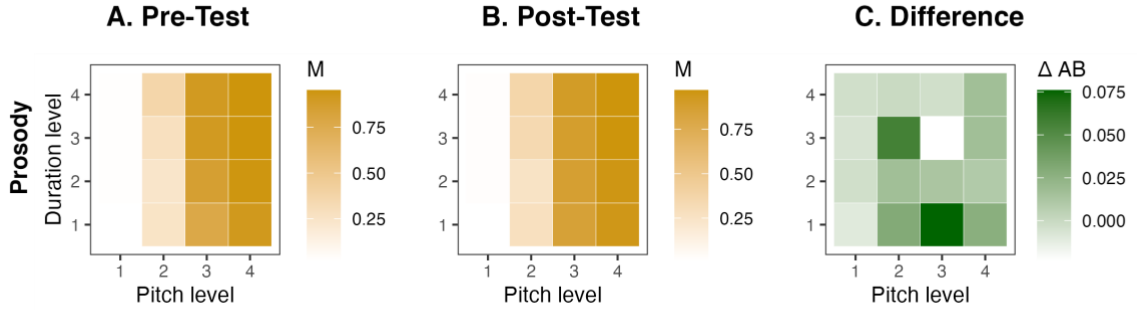
Time x Group	.100	.130	.772	.440
Time x Pitch	-1.618	.296	-5.464	**<.001**
Group x Pitch	1.728	.410	4.213	**<.001**
Time x Duration	-.073	.168	-.436	.663
Group x Duration	-.253	.194	-1.301	.193
Pitch x Duration	5.319	.428	12.432	**<.001**
Time x Group x Pitch	.433	.497	.871	.384
Time x Group x Duration	-.191	.256	-.747	.455
Time x Pitch x Duration	-1.907	.536	-3.561	**<.001**
Group x Pitch x Duration	-2.095	.708	-2.961	**.003**
Time x Group x Pitch x Duration	.817	.915	.893	.372

Linguistic focus

Results from the regression model (Table 24 and Figure 18) show that when categorizing linguistic focus stimuli, participants' decisions were influenced by both acoustic cues pitch (β=5.35, p<.001) and duration (β=.34, p<.001). The main effect of testing time (β=.25, p<.001) reflects a change in the bias towards one of the emphasized words over time.

A two-way interaction between session time and pitch level (β=.82, p<.001) suggests more pitch reliance at post-test, whereas a two-way interaction between training type and pitch level (β=.44, p=.035) stronger reliance on pitch in the Prosody training group. A three-way interaction between session and pitch and duration level (β=-.95, p=.024) suggests that at post-test, participants were less likely to integrate both dimensions during categorization. I observed no effects of training over time on any of the acoustic dimensions (p>.05), as evidenced by lack of the significant interaction between time, group and any of the acoustic dimensions.

A



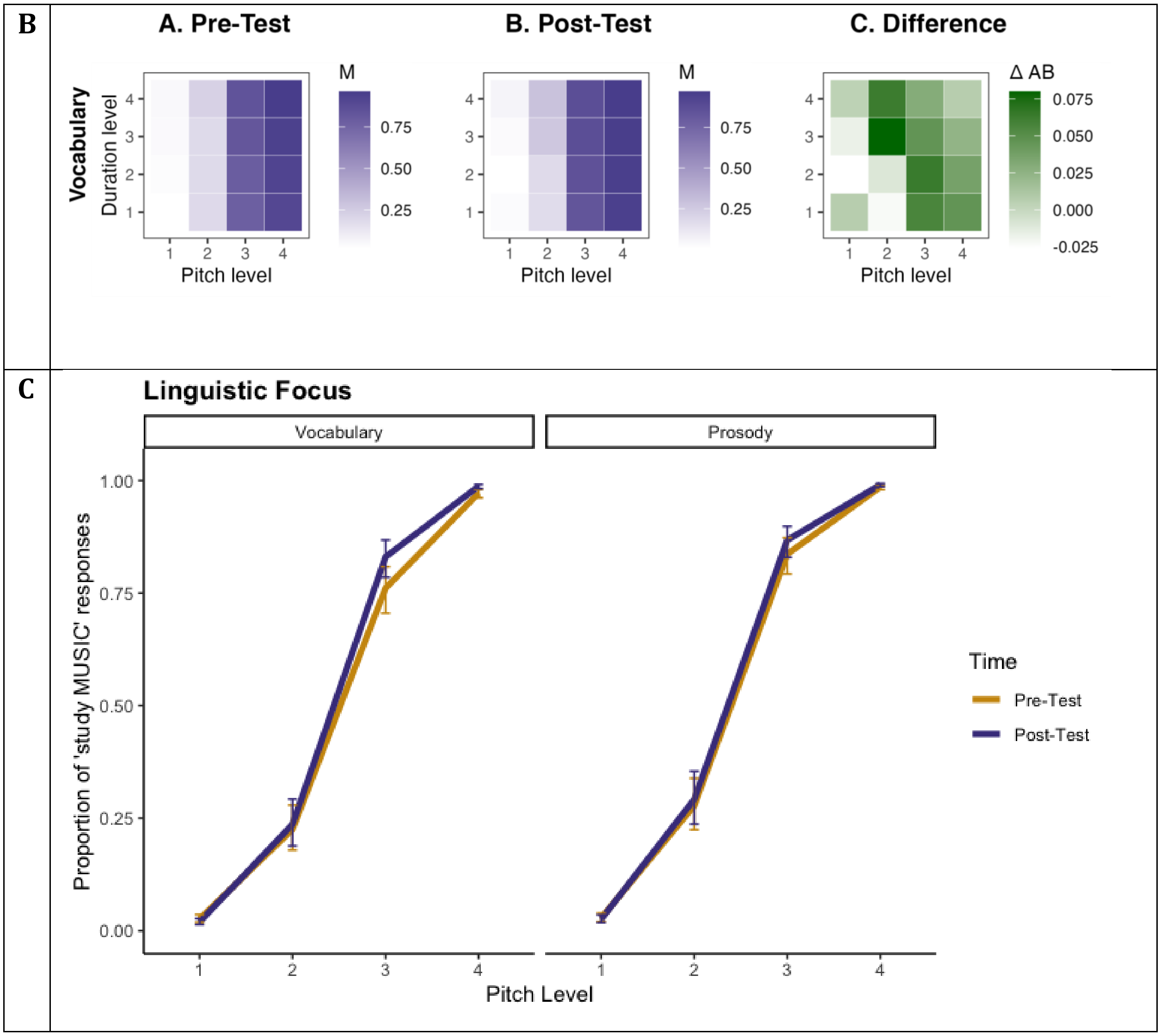
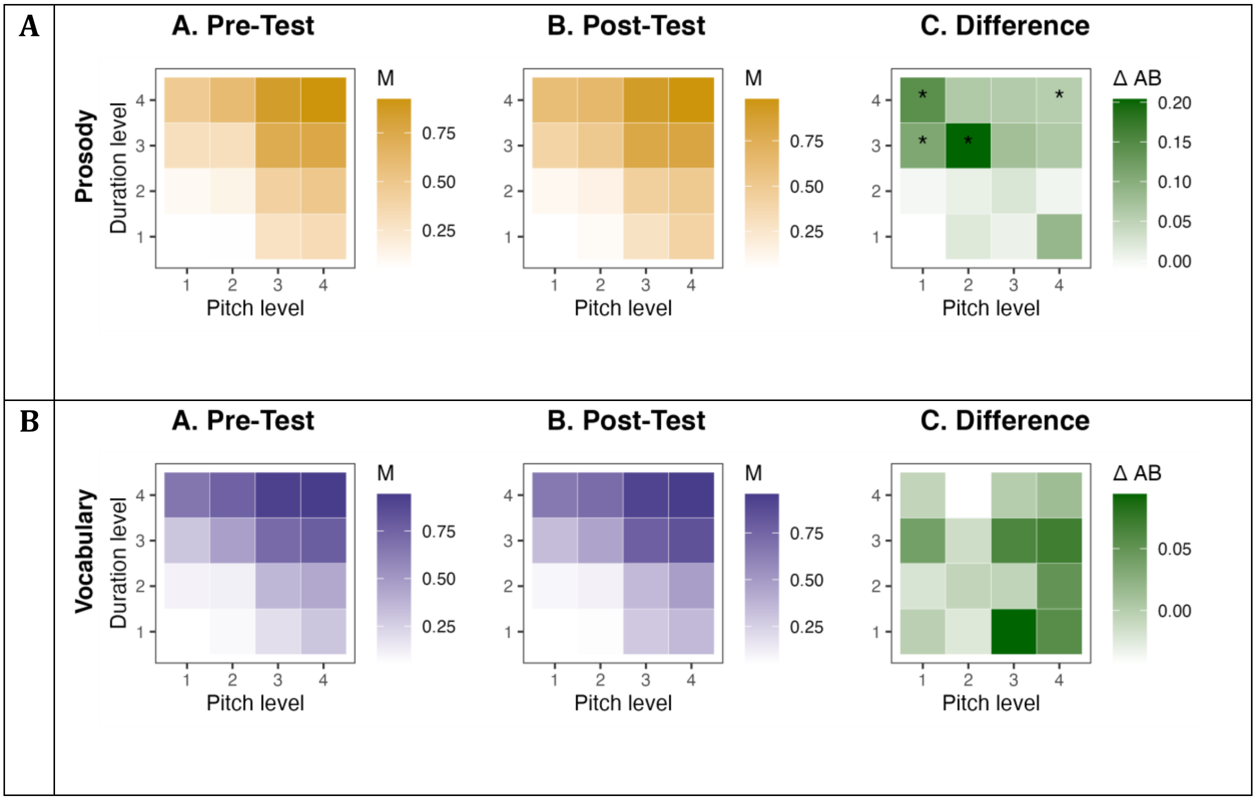


Figure 18. Linguistic focus categorization responses patterns. (18AB, ABC) Mean responses patterns at Pre-Test and Post-Test and the difference between the two sessions plotted separately for Prosody and Vocabulary training groups. (18C) Predicted proportion of "study MUSIC" vs "STUDY music" responses for Prosody and Vocabulary training groups at Pre-Test and Post-Test. Responses averaged across participants; error bars – 95% CI.

Phrase boundary

Results from logistic regression (Table 24 and Figure 19) show that participants' categorization of phrase boundary stimuli was influenced by both acoustic cues pitch (β=1.65, p<.001) and duration (β=2.90, p<.001). Several two-way interactions emerged as significant. An interaction between session time and training group (β=.22, p=.004) points to changes in bias towards one response to the other from pre-test to post-test in the Prosody group when compared to the Vocabulary training group. An interaction between session and pitch level (β=.26, p=.022) suggests stronger reliance on pitch at post-test, whereas interactions between training group and pitch level (β=.23, p=.043) and duration level (β=.65, p<.001) suggest stronger reliance on pitch and weaker on duration by Prosody group across testing sessions. A three-way interaction between testing session, training group and duration level (β=.42) suggests differences in how participants from different training groups use duration at pre-test and post-test. I followed up on this interaction with two separate regression models for each training group predicting categorization performance with time and duration level as predictors. This post-hoc analysis confirmed that participants from Prosody group relied more on duration after the training (β=.30, p=.002). No changes in duration use were observed for Vocabulary group(p>.05).



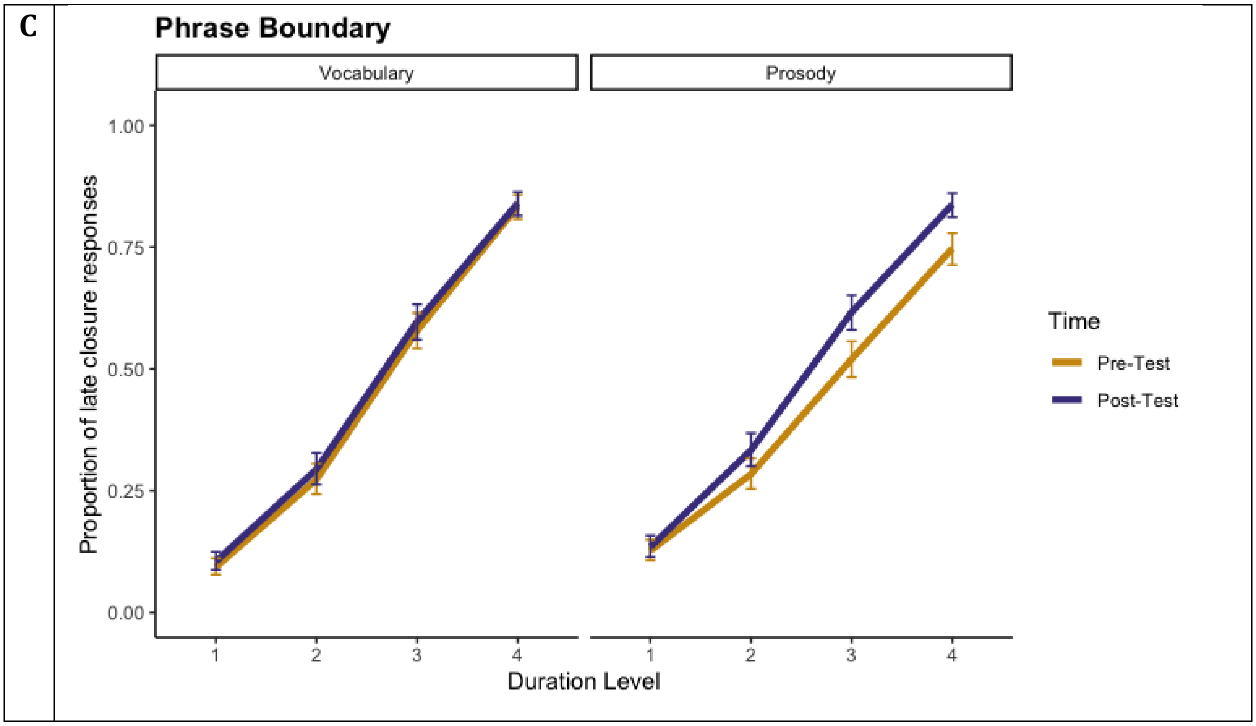


Figure 19. Phrase boundary categorization responses patterns. (19AB, ABC) Mean responses patterns at pre-test, post-test and the difference between the two sessions plotted separately for Prosody and Vocabulary training groups. All differences marked with asterisks were significant as shown by Mann-Whitney U tests with FDR correction for multiple comparisons. (19C) Predicted proportion of late vs early closure responses for Prosody and Vocabulary training groups at Pre-Test and Post-Test. Responses averaged across participants; error bars – 95%CI.

Lexical stress

Results from logistic regression (Table 24 and Figure 20) indicated that during categorization of lexical stress stimuli, participants were influenced by both acoustic features pitch (β=5.69, p<.001) and duration (β=.49, p<.001). The main effect of time (β=-.18, p=.013) points to changes from pre-test to post-test. A two-way interaction between testing session and training group (β=.27, p=.006) suggests changes from pre-test to post-test in Prosody group compared to Vocabulary training group. A three-way interaction between session, training group and pitch level (β=-1.29, p<.001) suggests differences in pitch use between the groups emerging over time. To interpret this interaction, I ran separate regression analyses for each training group with time and pitch level as predictors and categorization responses as outcome measure. The post hoc analyses revealed an increase in pitch use in Vocabulary group (β =.96, p=.001), but no change in pitch use in Prosody group (p>.05).

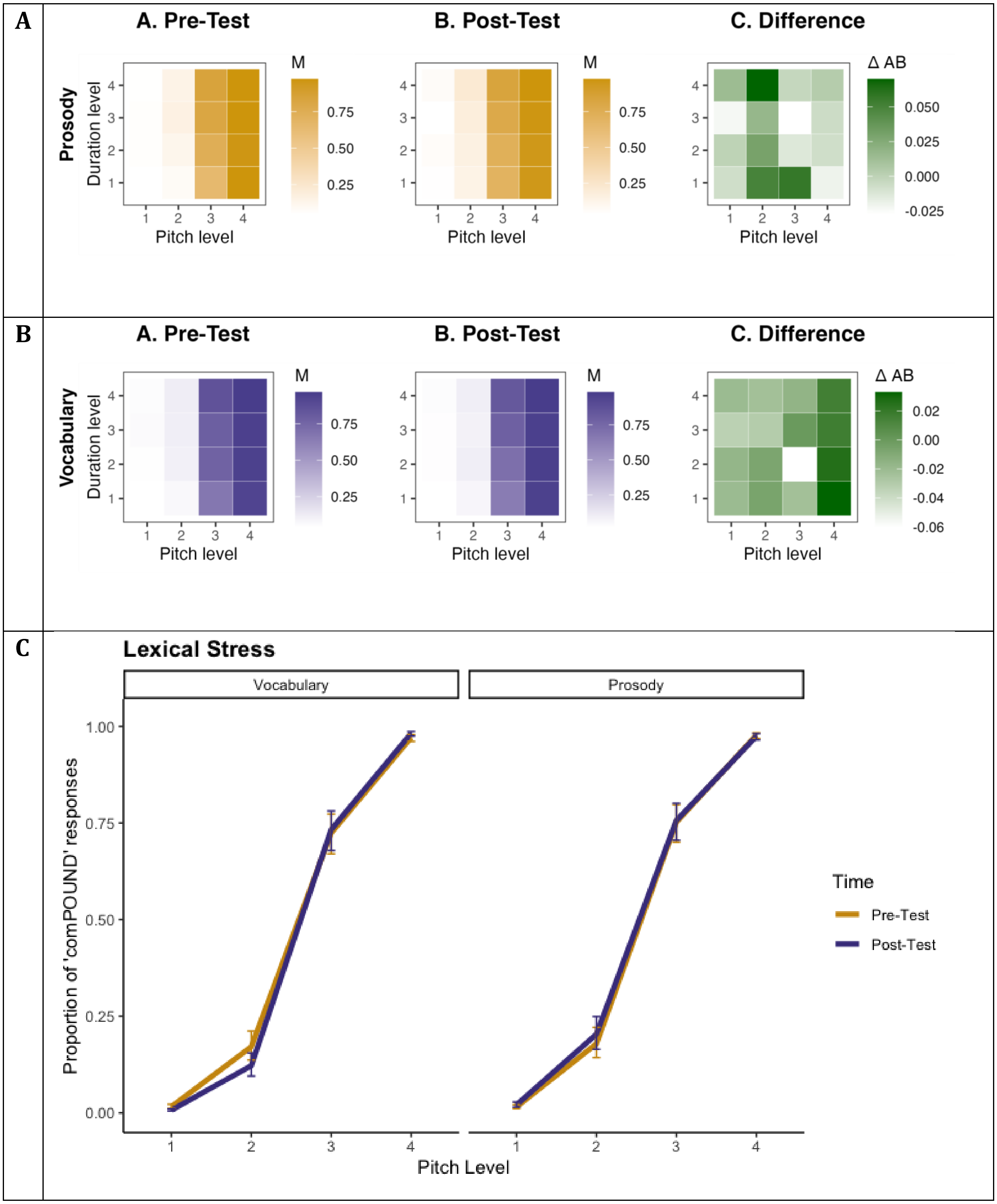
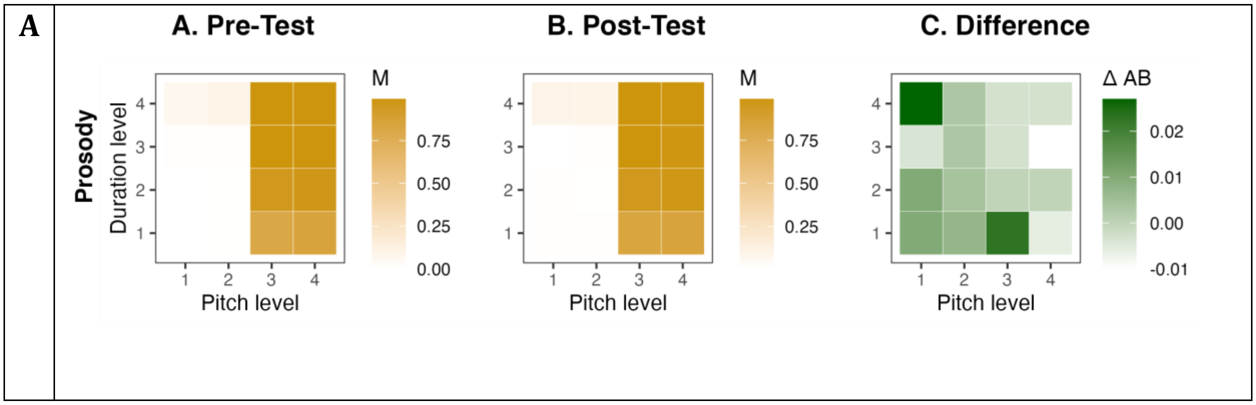


Figure 20. Lexical stress categorization responses patterns. (20A, ABC) Mean responses patterns at pre-test, post-test and the difference between the two sessions plotted separately for Prosody and Vocabulary training groups. (20C) Predicted proportion of "comPOUND" vs "COMpound" responses for Prosody and Vocabulary training groups at Pre-Test and Post-Test. Responses averaged across participants; error bars – 95%CI.

Musical beats

Results from logistic regression (Table 24 and Figure 21) demonstrated that participants' categorization of musical beats was influenced by both acoustic features pitch (β=8.70, p<.001) and duration (β=1.92, p<.011), as well as the combination of the two cues (β=5.32, p<.001). A two-way interaction between session time and pitch level (β=-1.62, p<.001) indicated less reliance on pitch during post-test, whereas a two-way interaction between training group and pitch level (β=1.73, p<.001) pointed to Prosody group participants being more reliant on pitch. A three-way interaction between session time, pitch level and duration level (β=-1.91, <.001) indicates that at post-test, participants were less likely to integrate across dimensions during categorization. Another three-way interaction between training group, pitch level and duration level (β=-2.10, p=.003) emphasizes that participants from Prosody group were less likely to integrate across the two dimensions. I observed no effects of training over time on any of the acoustic dimensions, as shown by non-significant interactions between time, group and both acoustic dimensions (p>.05).



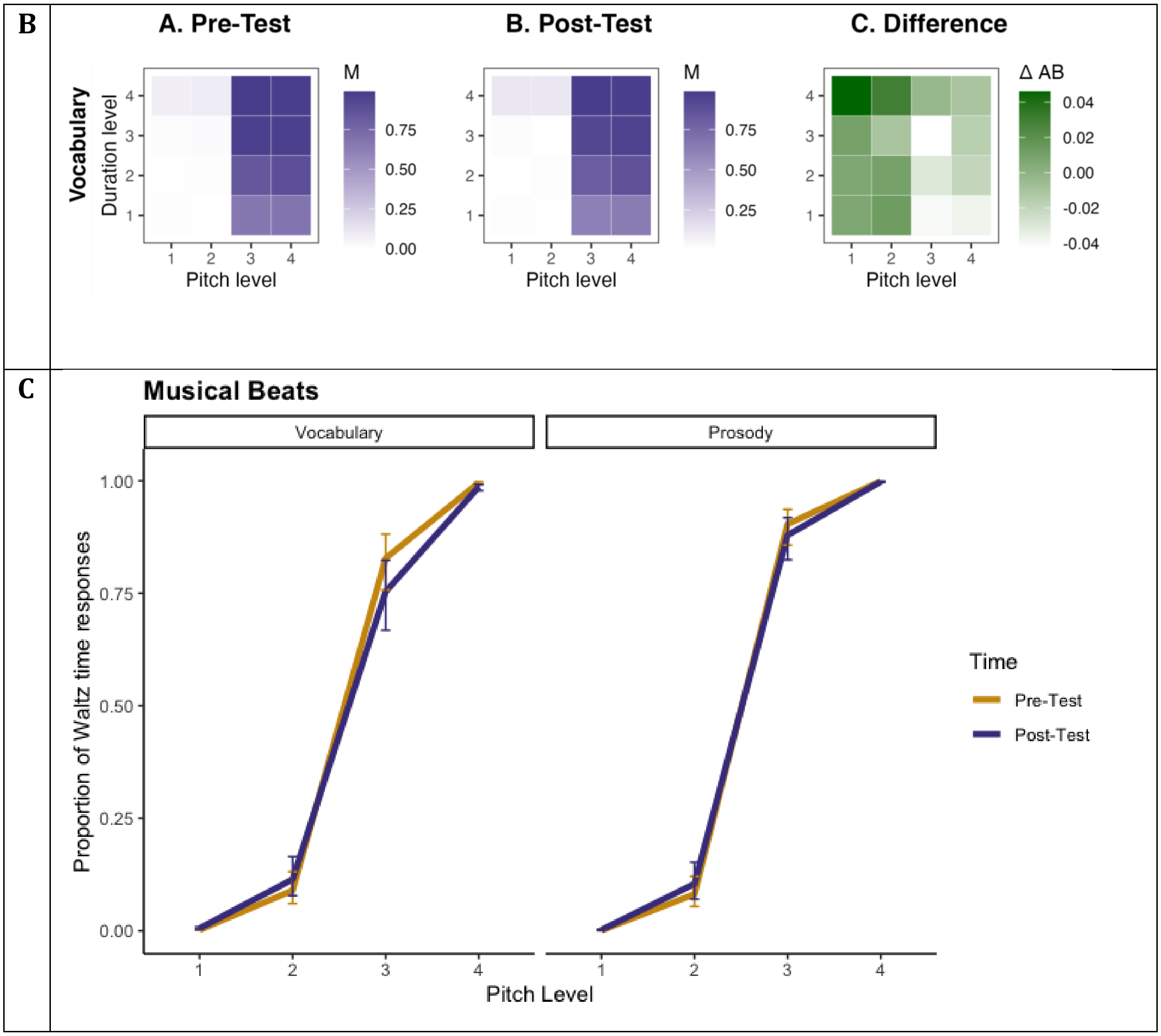


Figure 21. Musical beats categorization responses patterns. (21AB, ABC) Mean responses patterns at pre-test, post-test and the difference between the two sessions plotted separately for Prosody and Vocabulary training groups. (21C) Predicted proportion of Waltz time vs March time responses for Prosody and Vocabulary groups at Pre-Test and Post-Test. Responses averaged across participants; error bars – 95%CI.

Long-term training effects

I used a similar approach to examine the long-term training effects. The only difference from the cue weighting models reported above is that I used two different models with different coding systems. Model 1 examined whether participants showed continued improvements from Pre-Test to Post-Test I, and then from Post-Test I to Post-Test II. Model

2 was built to answer whether participants' performance returned to the baseline (i.e., to Pre-Test level) or if training effects were retained over time (i.e., comparison between Pre-Test and Post-Test II). For Model 1, I used the contr.sdif function from MASS package (Ripley et al., 2023) to estimate successive differences between Pre-Test (Time I) and Post-Test I (Time II) and then between Post-Test I (Time II) and Post-Test II after 6 months (Time III). In Model 2, the categorical variable representing the testing session (Pre-Test, Post-Test I and Post-Test II) was treatment coded so that both post-test sessions were compared to the baseline performance at Pre-Test (i.e., Time I to Time II and Time I to Time III). The main interactions of interest that answer my research questions are the three-way interactions between time, group and pitch (estimating the differences in pitch use between groups and across time) and between time, group and duration (estimating the differences in duration use between groups and across time) for comparisons between Time I and Time II and then Time II an Time III in Model 1, and the additional comparisons for Time I and III in Model 2. The main effects of interest were the three-way interactions between the training group, testing time and use of pitch or duration. All results of the mixed-effects logistic regression models are presented in Table 25.

Table 25. Summary of effects in mixed-effects logistic regression models for categorization tasks at Pre-Test, Post-Test I and Post-Test II.

	Model 1				***Model 2***			
Predictor	***Estimate***	***SE***	***z***	***p***	***Estimate***	***SE***	***z***	***p***
Linguistic Focus								
Intercept	.019	.129	.148	.883	-.162	.135	-1.202	.230
Time (I⟶II)	.307	.076	4.019	**<.001**	.307	.076	4.019	**<.001**
Time (II⟶III)	-.070	.079	-.885	.376	--	--	--	--
Time (I⟶III)	--	--	--	--	.237	.075	3.158	**.002**
Group (Vocabulary)	.317	.183	1.736	.083	.527	.193	2.736	**.006**
Pitch	5.331	.100	53.285	**<.001**	4.751	.147	32.368	**<.001**
Duration	.360	.063	5.745	**<.001**	.329	.103	3.201	**.001**
Time (I⟶II) x Group	-.261	.112	-2.330	**.020**	-.261	.112	-2.329	**.020**
Time (II⟶III) x Group	-.109	.114	-.957	.338	--	--	--	--
Time (I⟶III) x Group	--	--	--	--	-.370	.111	-3.336	**<.001**
Time (I⟶II) x Pitch	1.021	.228	4.473	**<.001**	1.021	.228	4.468	**<.001**
Time (II⟶III) x Pitch	-.300	.244	-1.231	.219	--	--	--	--
Time (I⟶III) x Pitch	--	--	--	--	.721	.221	3.269	**.001**
Group x Pitch	.579	.150	3.873	**<.001**	.992	.235	4.224	**<.001**
Time (I⟶II) x Duration	.095	.152	.625	.532	.095	.152	.625	.532
Time (II⟶III) x Duration	-.099	.157	-.633	.527	--	--	--	--
Time (I⟶III) x Duration	--	--	--	--	-.004	.150	-.027	.979
Group x Duration	.093	.092	1.011	.312	.201	.154	1.308	.191
Pitch x Duration	.122	.188	.650	.516	.417	.286	1.456	.146
Time (I⟶II) x Group x Pitch	-.829	.347	-2.392	**.017**	-.829	.347	-2.391	**.017**
Time (II⟶III) x Group x Pitch	.421	.362	1.163	.245	--	--	--	--
Time (I⟶III) x Group x Pitch	--	--	--	--	-.408	.343	-1.188	.235
Time (I⟶II) x Group x Duration	-.147	.224	-.657	.511	-.147	.224	-.657	.511
Time (II⟶III) x Group x Duration	-.032	.227	-.141	.887	--	--	--	--
Time (I⟶III) x Group x Duration	--	--	--	**--**	-.179	.222	-.809	.418
Time (I⟶II) x Pitch x Duration	-1.178	.451	-2.613	**.009**	-1.178	.455	-2.952	**.009**
Time (II⟶III) x Pitch x Duration	1.472	.482	3.056	**.002**	--	--	--	--
Time (I⟶III) x Pitch x Duration	--	--	--	--	.294	.438	.671	.502
Group x Pitch x Duration	.150	.285	.525	.560	.253	.459	.550	.582

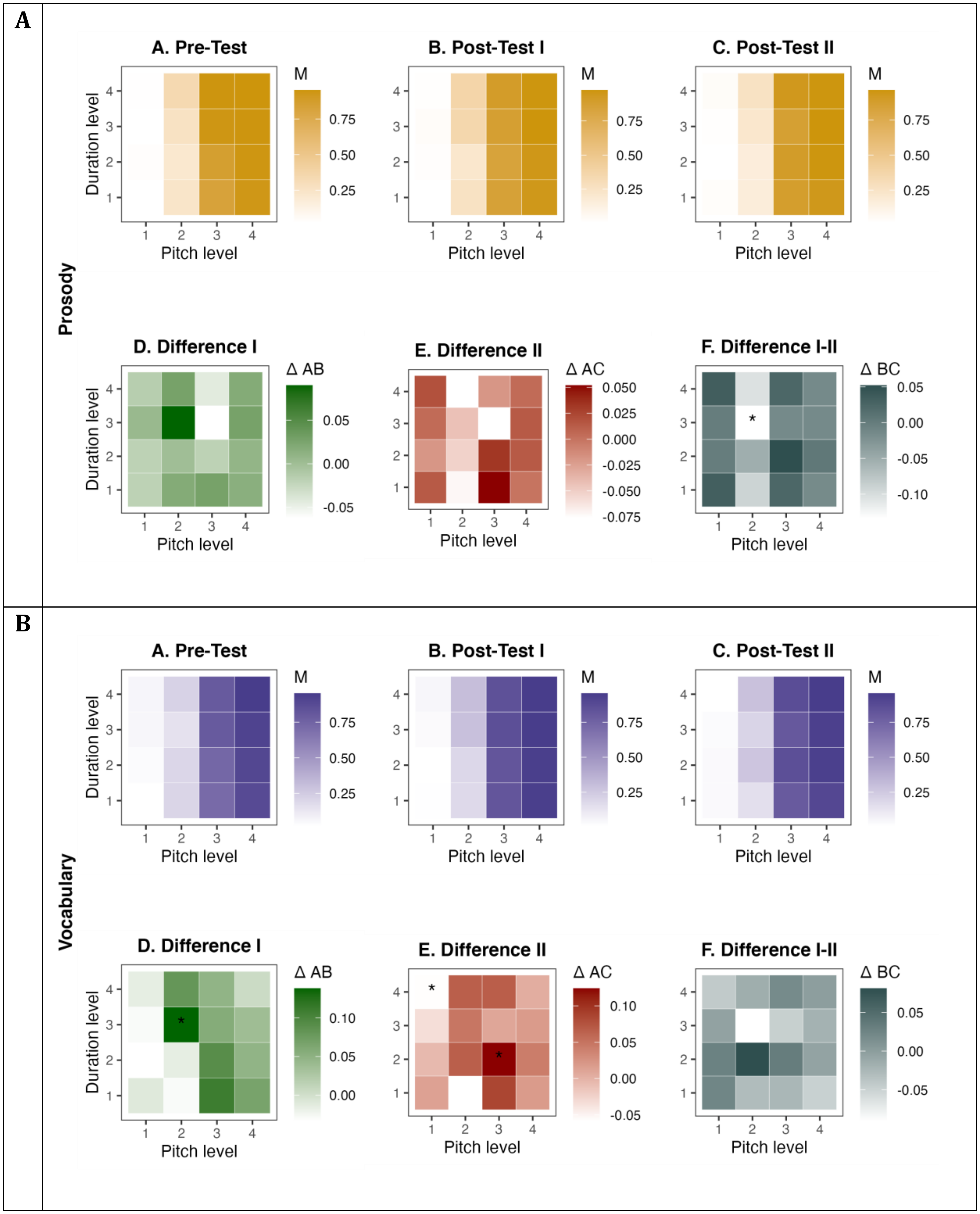
Time (I⟶II) x Group x Pitch x Duration	.683	.684	.999	.318	.684	.689	.993	.321
Time (II⟶III) x Group x Pitch x Duration	-1.677	.717	-2.339	**.019**	--	--	--	--
Time (I⟶III) x Group x Pitch x Duration	--	--	--	--	-.994	.683	-1.455	.146
Phrase Boundary								
Intercept	-.236	.075	-3.146	**.002**	-.283	.083	-3.401	**<.001**
Time (I⟶II)	.026	.064	.410	.682	.026	.064	.409	.682
Time (II⟶III)	.090	.065	1.380	.168	--	--	--	--
Time (I⟶III)	--	--	--	--	.116	.065	1.786	.074
Group (Vocabulary)	-.162	.106	-1.523	.128	-.139	.118	-1.179	.238
Pitch	1.617	.054	30.088	**<.001**	1.364	.091	15.056	**<.001**
Duration	2.713	.058	46.471	**<.001**	2.738	.100	27.453	**<.001**
Time (I⟶II) x Group	.120	.091	1.317	.188	.120	.091	1.316	.188
Time (II⟶III) x Group	-.308	.094	-3.289	**.001**	--	--	--	--
Time (I⟶III) x Group	--	--	--	--	-.188	.093	-2.021	**.043**
Time (I⟶II) x Pitch	.304	.129	2.353	**.019**	.304	.129	2.353	**.019**
Time (II⟶III) x Pitch	.152	.133	1.146	.252	--	--	--	--
Time (I⟶III) x Pitch	--	--	--	**--**	.456	.132	3.458	**<.001**
Group x Pitch	.319	.078	4.102	**<.001**	.510	.131	3.899	**<.001**
Time (I⟶II) x Duration	-.177	.140	-1.264	.206	-.177	.140	-1.265	.206
Time (II⟶III) x Duration	.277	.143	1.939	.053	--	--	--	--
Time (I⟶III) x Duration	--	--	--	--	.100	.144	.696	.486
Group x Duration	-.133	.083	-1.603	.109	-.334	.140	-2.387	**.017**
Pitch x Duration	-.190	.115	-1.661	.097	-.032	.197	-.162	.871
Time (I⟶II) x Group x Pitch	-.100	.187	-.534	.594	-.100	.187	-.533	.594
Time (II⟶III) x Group x Pitch	-.375	.192	-1.957	.050	--	--	--	--
Time (I⟶III) x Group x Pitch	--	--	--	--	-.475	.190	-2.507	**.012**
Time (I⟶II) x Group x Duration	.353	.198	1.780	.075	.353	.198	1.781	.075
Time (II⟶III) x Group x Duration	-.103	.204	-.508	.611	--	--	--	--
Time (I⟶III) x Group x Duration	--	--	--	--	.249	.203	1.231	.218
Time (I⟶II) x Pitch x Duration	-.186	.276	-.676	.499	-.186	.276	-.675	.499
Time (II⟶III) x Pitch x Duration	-.102	.280	-.364	.716	--	--	--	--
Time (I⟶III) x Pitch x Duration	--	--	--	--	-.289	.283	-1.020	.308
Group x Pitch x Duration	-.064	.162	-.391	.696	.180	.277	.648	.517
Time (I⟶II) x Group x Pitch x Duration	-.320	.391	-.817	.414	-.320	.392	-.814	.415
Time (II⟶III) x Group x Pitch x Duration	-.091	.400	-.228	.820	--	--	--	--
Time (I⟶III) x Group x Pitch x Duration	--	--	--	--	-.410	.400	-1.026	.305

Lexical Stress								
Intercept	-.483	.163	-2.966	**.003**	-.372	.168	-2.217	**.027**
Time (I⟶II)	-.092	.080	-1.155	.248	-.092	.080	-1.155	.248
Time (II⟶III)	-.149	.091	-1.626	.104	--	--	--	--
Time (I⟶III)	--	--	--	--	-.241	.088	-2.751	**.006**
Group (Vocabulary)	.181	.230	.784	.433	.096	.239	.402	.698
Pitch	6.223	.123	50.711	**<.001**	4.992	.156	32.066	**<.001**
Duration	.505	.071	7.149	**<.001**	.414	.107	3.866	**<.001**
Time (I⟶II) x Group	.149	.119	1.255	.210	.149	.119	1.255	.210
Time (II⟶III) x Group	-.046	.126	-.364	.716	--	--	--	--
Time (I⟶III) x Group	--	--	--	**--**	.104	.122	.848	.396
Time (I⟶II) x Pitch	1.097	.245	4.480	**<.001**	1.096	.245	4.477	**<.001**
Time (II⟶III) x Pitch	1.518	.317	4.793	**<.001**	--	--	--	--
Time (I⟶III) x Pitch	--	--	--	**--**	2.614	.296	8.847	**<.001**
Group x Pitch	.431	.179	2.413	**.016**	1.806	.266	6.785	**<.001**
Time (I⟶II) x Duration	.004	.160	.025	.980	.004	.160	.025	.980
Time (II⟶III) x Duration	.267	.183	1.461	.144	--	--	--	--
Time (I⟶III) x Duration	--	--	--	--	.271	.175	1.545	.122
Group x Duration	.035	.100	.347	.729	.121	.163	.737	.461
Pitch x Duration	.722	.232	3.118	**.002**	.338	.303	1.114	.265
Time (I⟶II) x Group x Pitch	-.804	.390	-2.063	**.039**	-.804	.390	-2.063	**.039**
Time (II⟶III) x Group x Pitch	-2.517	.427	-5.897	**<.001**	--	--	--	--
Time (I⟶III) x Group x Pitch	--	--	--	--	-3.321	.405	-8.189	**<.001**
Time (I⟶II) x Group x Duration	.092	.238	.386	.700	.092	.238	.385	.700
Time (II⟶III) x Group x Duration	0.441	.251	-1.754	.079	--	--	--	--
Time (I⟶III) x Group x Duration	--	--	--	--	-.349	.245	-1.427	.153
Time (I⟶II) x Pitch x Duration	.118	.484	.244	.807	.118	.485	.243	.808
Time (II⟶III) x Pitch x Duration	.917	.621	1.476	.140	--	--	--	--
Time (I⟶III) x Pitch x Duration	--	--	--	--	1.035	.578	1.791	.073
Group x Pitch x Duration	-.362	.330	-1.097	.273	-.026	.512	-.050	.960
Time (I⟶II) x Group x Pitch x Duration	-.003	.768	-.004	.997	-.03	.769	-.004	.997
Time (II⟶III) x Group x Pitch x Duration	-1.003	.840	-1.194	.232	--	--	--	--
Time (I⟶III) x Group x Pitch x Duration	--	--	--	--	-1.007	.796	-1.264	.206
Musical Beats								
Intercept	-.530	.269	-1.972	**.049**	-.515	.276	-1.865	.062
Time (I⟶II)	-.071	.101	-.698	.485	-.071	.101	-.698	.485

Time (II⟶III)	.094	.099	.958	.338	--	--	--	--
Time (I⟶III)	--	--	--	--	.024	.104	.229	.819
Group (Vocabulary)	.411	.381	1.079	.281	.464	.394	1.177	.239
Pitch	8.101	.180	45.060	**<.001**	8.793	.293	29.999	**<.001**
Duration	1.775	.087	20.401	**<.001**	1.873	.152	12.292	**<.001**
Time (I⟶II) x Group	-.033	.160	-.207	.836	-.033	.160	-.207	.836
Time (II⟶III) x Group	-.094	.146	-.640	.522	--	--	--	--
Time (I⟶III) x Group	--	--	--	**--**	-.127	.159	-.799	.424
Time (I⟶II) x Pitch	-1.539	.345	-4.461	**<.001**	-1.539	.344	-4.473	**<.001**
Time (II⟶III) x Pitch	1.002	.334	3.005	**.003**	--	--	--	--
Time (I⟶III) x Pitch	--	--	--	**--**	-.537	.368	-1.459	**<.001**
Group x Pitch	1.455	.279	5.212	**<.001**	2.210	.509	4.345	**<.001**
Time (I⟶II) x Duration	-.289	.200	-1.446	.148	-.289	.200	-1.148	.148
Time (II⟶III) x Duration	.284	.195	1.458	.145	--	--	--	--
Time (I⟶III) x Duration	--	--	--	**--**	-.005	.206	-.024	.981
Group x Duration	-.301	.129	-2.326	**.020**	-.216	.240	-.899	.369
Pitch x Duration	4.590	.284	16.187	**<.001**	4.561	.505	9.037	**<.001**
Time (I⟶II) x Group x Pitch	.014	.621	.022	.983	.014	.622	.022	.983
Time (II⟶III) x Group x Pitch	-2.290	.539	-4.247	**<.001**	--	--	--	--
Time (I⟶III) x Group x Pitch	--	--	--	--	-2.276	.610	-3.729	**<.001**
Time (I⟶II) x Group x Duration	-.017	.315	-.054	.957	-.017	.315	-.054	.957
Time (II⟶III) x Group x Duration	-.222	.289	-.769	.442	--	--	--	--
Time (I⟶III) x Group x Duration	--	--	--	--	-.239	.311	-.769	.442
Time (I⟶II) x Pitch x Duration	-.759	.643	-1.180	.238	-.759	.542	-1.182	.237
Time (II⟶III) x Pitch x Duration	1.606	.620	2.592	**.010**	--	--	--	--
Time (I⟶III) x Pitch x Duration	--	--	--	--	.847	.676	1.252	.211
Group x Pitch x Duration	-.069	.454	-.152	.880	1.830	.870	2.103	**.035**
Time (I⟶II) x Group x Pitch x Duration	-1.514	1.114	-1.359	.174	-1.515	1.123	-1.349	.177
Time (II⟶III) x Group x Pitch x Duration	-2.666	.995	-2.680	**.007**	--	--	--	--
Time (I⟶III) x Group x Pitch x Duration	--	--	--	**--**	-4.181	1.100	-3.800	**<.001**

Linguistic focus

Results from logistic regression (Table 25 and Figure 22, Model 1) demonstrated that participants' categorization of linguistic focus was influenced by both acoustic features pitch (β=5.33, p<.001) and duration (β=.36, p<.011). The main effect of session suggests a change in cue weighting from Pre-Test to Post-Test I (β=.31, p<.001). There was an overall increase in pitch use from Pre-Test to Post-Test I (β=1.02, p<.001). A two-way interaction between the training group and pitch level (β =.58, p<.001) suggests more reliance on pitch in the Prosody training group. A significant three-way interaction between time, group and pitch level (β=-.83, p=.017) suggests emerging differences between groups over time in their use of pitch. A four-way interaction between session, training group and both dimensions (β=-.68, p=.019) suggests that the strength of the change in the interaction between pitch and duration cues over time differed between the groups. To follow up on the significant three-way interaction, I built separate regression models for each training group to predict participants' categorization responses with time and pitch level as predictors. This post hoc analysis revealed that Vocabulary group significantly increased in their pitch reliance from Pre-Test to Post-Test I (β=.983, p<.001), but no further changes were observed at Post-Test II (p>.05). No changes in pitch use were observed for the Prosody group. Model 2 estimating the additional comparison between Pre-Test and Post-Test II showed the- overall increase in pitch use from Pre-Test to Post-Test II (β=.721, p=.001), but no significant differences between the training groups at Pre-Test and Post-Test II were found for either pitch or duration use (p>.05).


A
Prosody
A. Pre-Test
B. Post-Test I
C. Post-Test II
M
0.75
0.50
0.25
Duration level
Pitch level
1
2
3
4
D. Difference I
Δ AB
0.05
0.00
-0.05
E. Difference II
Δ AC
0.050
0.025
0.000
-0.025
-0.050
-0.075
F. Difference I-II
Δ BC
0.05
0.00
-0.05
-0.10
*
B
Vocabulary
A. Pre-Test
B. Post-Test I
C. Post-Test II
M
0.75
0.50
0.25
D. Difference I
Δ AB
0.10
0.05
0.00
E. Difference II
Δ AC
0.10
0.05
0.00
-0.05
F. Difference I-II
Δ BC
0.05
0.00
-0.05

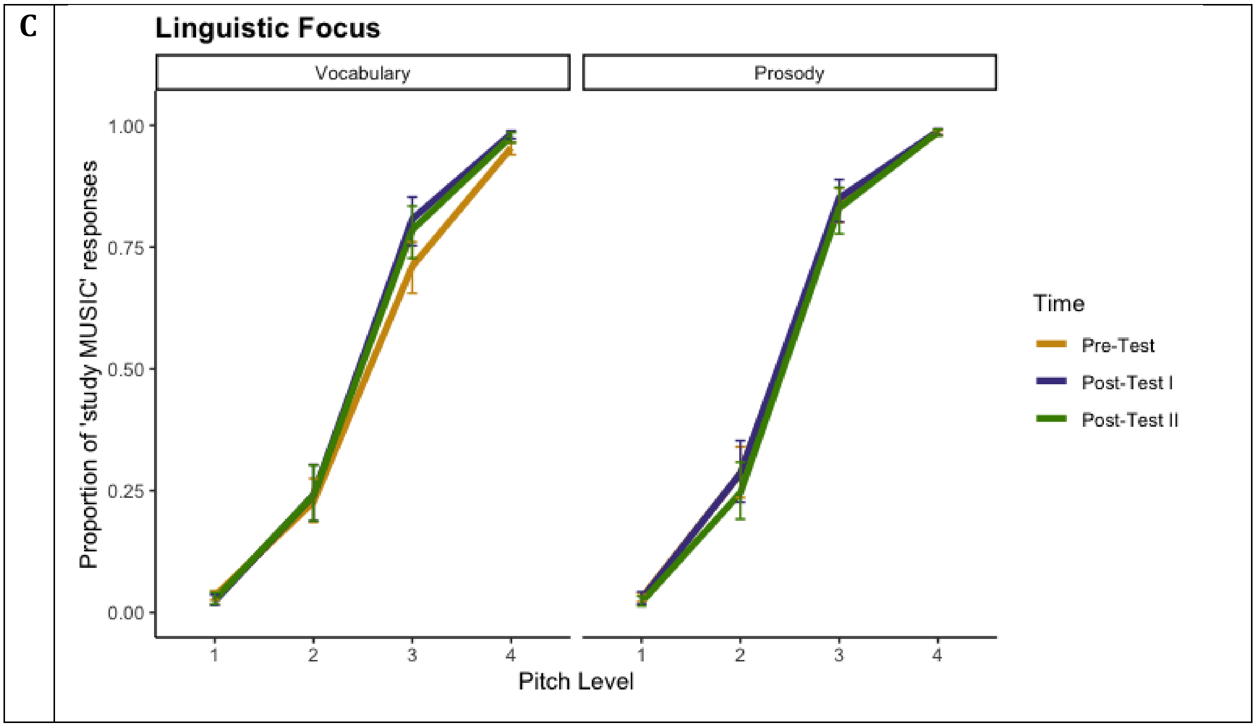
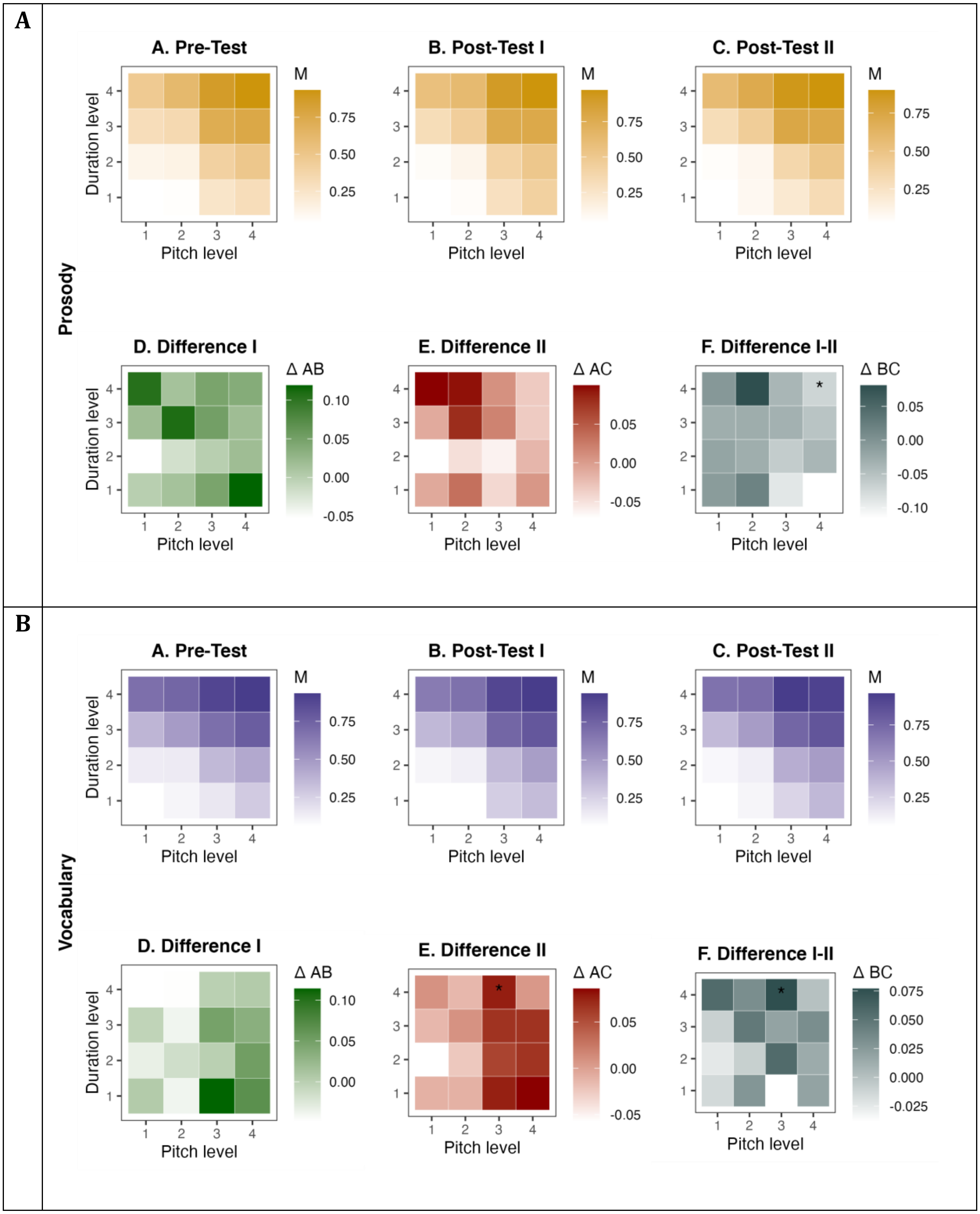


Figure 22. Linguistic focus categorization responses patterns. (22AB, A-F) Mean responses patterns at Pre-Test, Post-Test I and Post-Test II and the differences between the sessions plotted separately for Prosody and Vocabulary training groups. All differences marked with asterisks were significant as shown by Mann-Whitney U tests with FDR correction for multiple comparisons. (22C) Predicted proportion of "study MUSIC" vs "STUDY music" responses for Prosody and Vocabulary training groups at Pre-Test, Post-Test I and Post-Test II. Responses averaged across participants; error bars – 95%CI.

Phrase boundary

Results from logistic regression (Table 25 and Figure 23) show that participants categorization of phrase boundaries was influenced by both acoustic features pitch (β=1.62, p<.001) and duration (β=2.72, p<.001). There seems to be an overall change in response bias among participants from Prosody training group from Post-Test I to Post-Test II (β=-.31, p=.001). I also observed an overall increase in pitch use from Pre-Test to Post-Test I (β=.31, p=.018) and stronger reliance on pitch in the Prosody Training group (β=.32, p<.001). A three-way interaction suggesting decrease in pitch use from Post-Test I to Post-Test II in Prosody training group (β=-.38, p=.05035) missed significance. However, Model 2 yielded a significant three-way interaction between time (Pre-Test vs Post-Test II), group and pitch use (β=-.475, p=.012). Post-hoc regression models for each training group revealed that this difference was driven by small, but systematic increase in pitch use by the Vocabulary group (Time I vs Time II, β=.256, p=.014; Time I vs Time III, β=.298, p=.004).



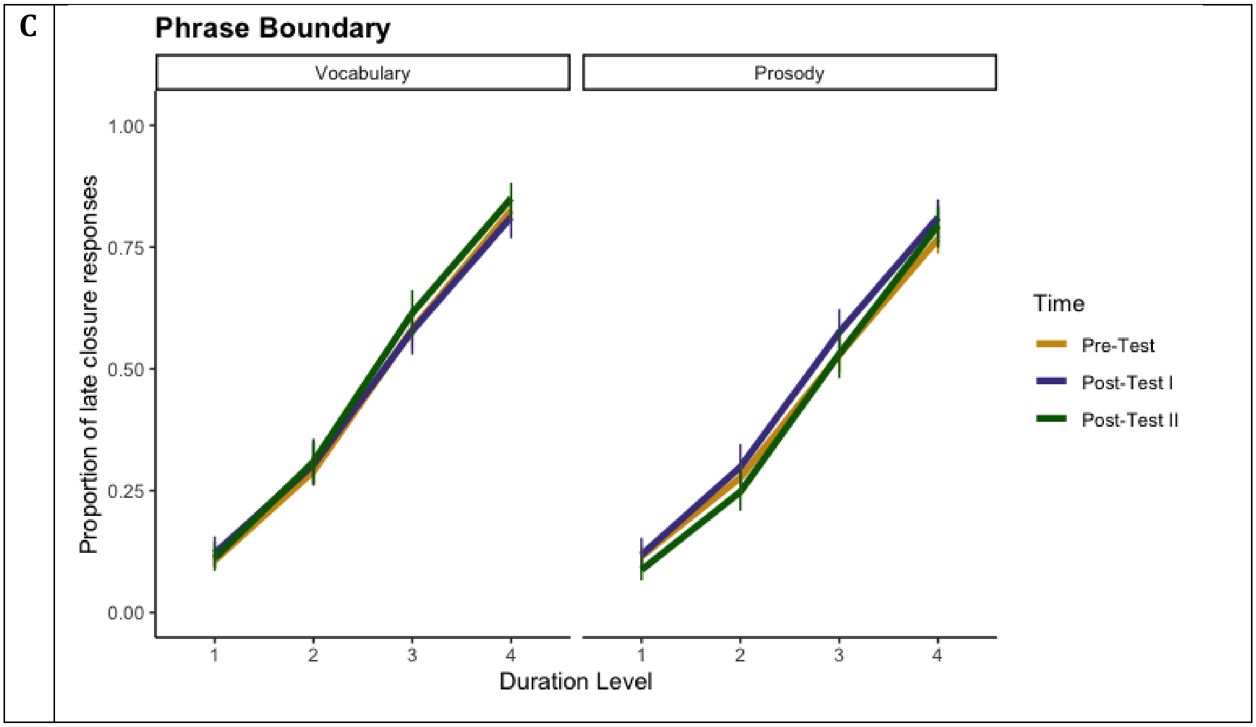
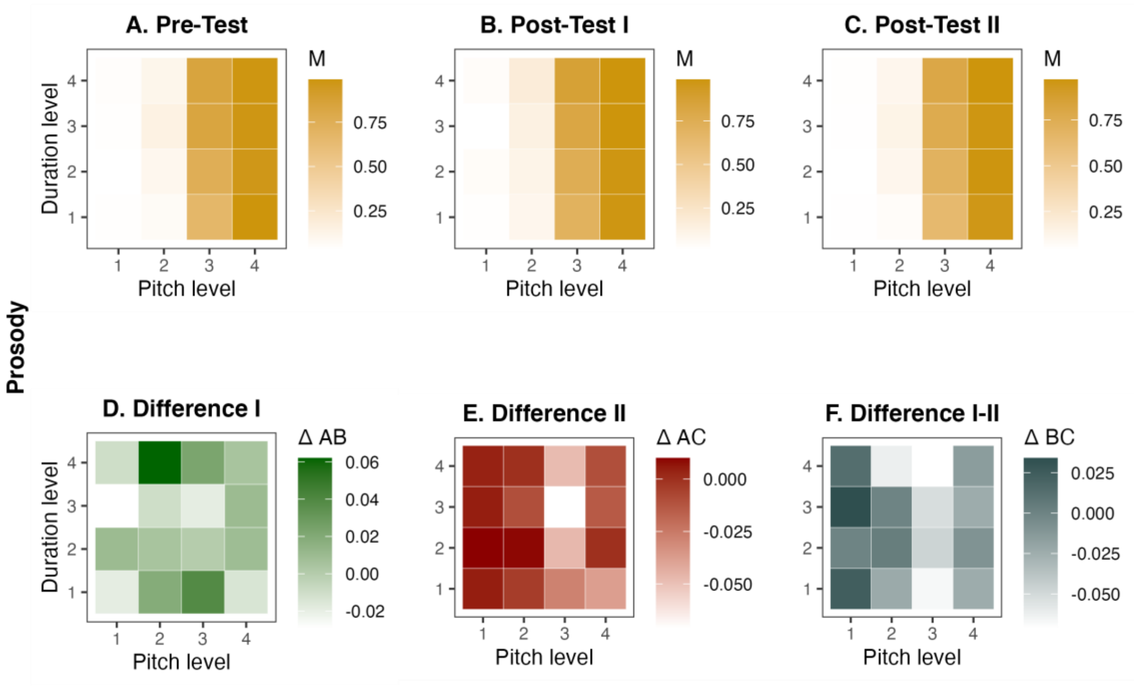


Figure 23. Phrase boundary categorization responses patterns. (23AB, A-F) Mean responses patterns at Pre-Test, Post-Test I and Post-Test II and the differences between the sessions plotted separately for Prosody and Vocabulary training groups. All differences marked with asterisks were significant as shown by Mann-Whitney U tests with FDR correction for multiple comparisons. (23C) Predicted proportion of early vs late closure responses for Prosody and Vocabulary training groups at Pre-Test, Post-Test I and Post-Test II. Responses averaged across participants; error bars – 95%CI.

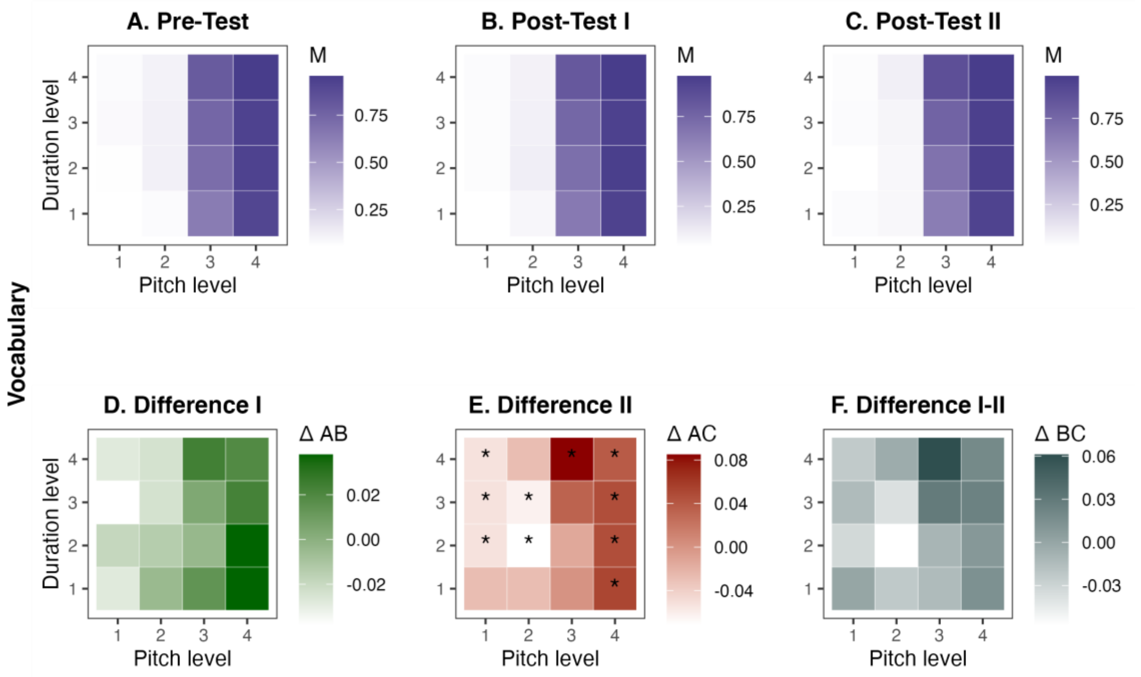
Lexical stress

Results from logistic regression (Table 25 and Figure 24) show that when categorizing lexical stress participants were influenced by both acoustic features, pitch (β=6.23, p<.001) and duration (β=0.51, p<.001). There was also a significant effect of interaction between pitch and duration (β=.72, p=.002), suggesting the interdependence of both types of information in categorizing lexical stress. I observed an overall increase in pitch reliance from Pre-Test to Post-Test I (β=1.10, p<.001) and from Post-Test I to Post-Test II (β=1.52, p<.001). The Prosody group relied more on pitch compared to the Vocabulary training group (β=.43, p=.016). I also found significant three-way interactions of training group and pitch use with time at Pre-Test vs Post-Test I (β=-.80, p=.039) and at Post-Test I vs Post-Test II (β=2.52, p<.001). To follow up on these effects, I ran two separate regressions for each training group and time and pitch levels as predictors. These post hoc analyses revealed systematic increase in pitch use in Vocabulary group from Pre-Test to Post-Test I (b=1.08, p<.001) and from Post-Test I to Post-Test II (β=1.36, p<.001). Additionally, there was a significant decrease in pitch reliance in Prosody group from Post-Test I to Post-Test II (β=.947, p<.001). Model 2 also yielded a significant three-way interaction of training group at Pre-Test vs Post-Test II (β=-3.321, p<.001). Post-hoc analyses confirmed systematic increase in pitch use in Vocabulary group (Pre-Test vs Post-Test II, β=2.437, p<.001), but also revealed decrease in pitch from Pre-Test to Post-Test II in prosody group (β=-.687, p=.012).

A



B



C

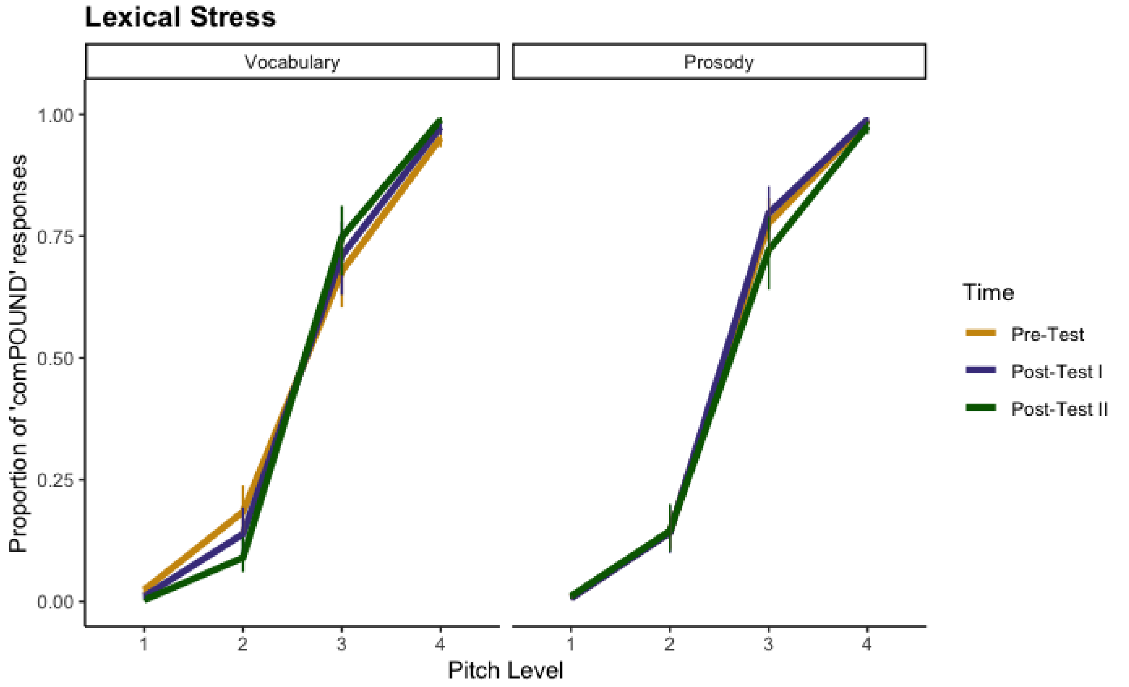


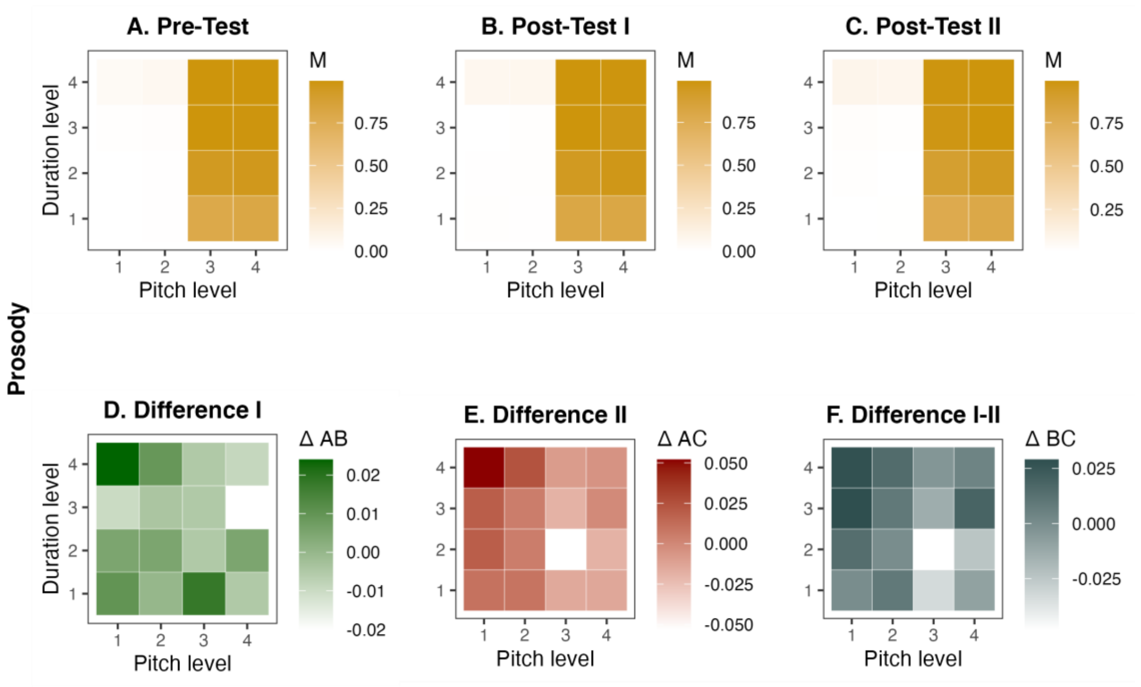
Figure 24. Lexical stress categorization responses patterns. (24AB, A-F) Mean responses patterns at Pre-Test, Post-Test I and Post-Test II and the differences between the sessions plotted separately for Prosody and Vocabulary training groups. All differences marked with asterisks were significant as shown by Mann-Whitney U tests with FDR correction for multiple comparisons. (24C) Predicted proportion of “comPOUND" vs “COMpound”

responses for Prosody and Vocabulary training groups at Pre-Test, Post-Test I and Post-Test II. Responses averaged across participants.; error bars – 95%CI.

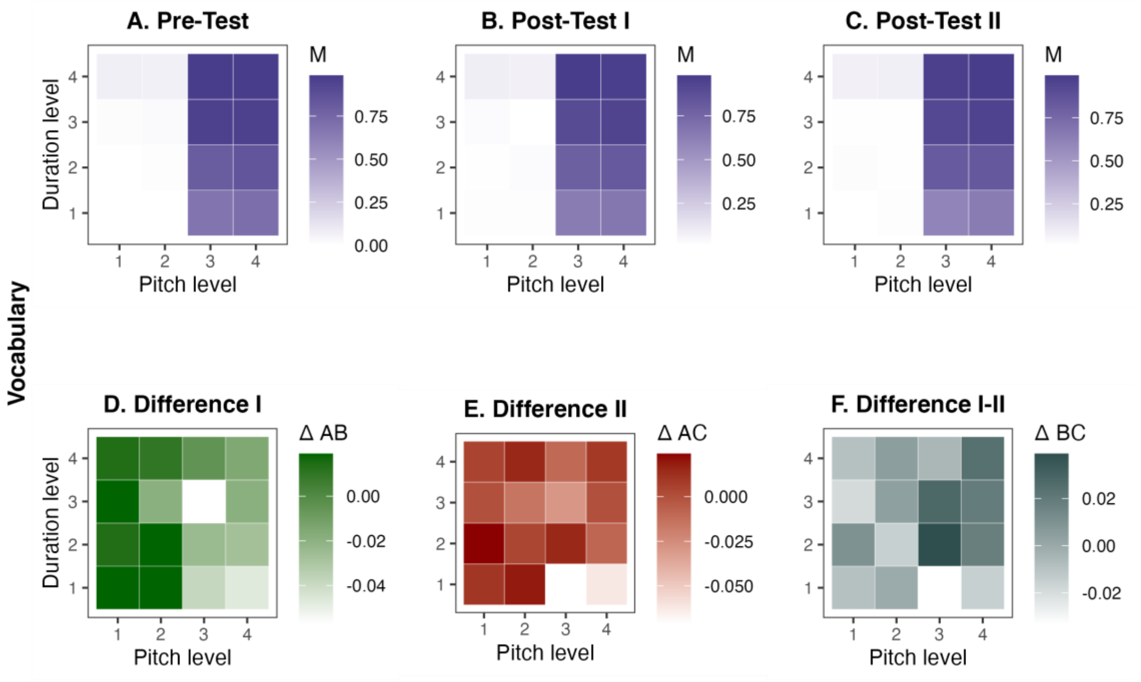
Musical beats

Results from logistic regression (Table 25 and Figure 25) demonstrated that participants' categorization of musical beats was influenced by both acoustic features pitch (β=8.10, p<.001) and duration (β=1.76, p<.001). There was also a significant effect of interaction between pitch and duration (β=4.59, p<.001), suggesting the interdependence of both types of information in categorizing musical beats. I observed an overall decrease in pitch reliance from Pre-Test to Post-Test I (β=-1.54, p<.001) and increase from Post-Test I to Post-Test II (β=1.00, p=.003). The Prosody group relied more on pitch (β=1.46, p<.001) and less on duration (β=-.30, p=.02) compared to the Vocabulary group. A significant three-way interaction between session (Post-Test II vs Post-Test III), training and pitch suggest differences in reliance on pitch between the training groups (β=-2.29, p<.001). A three-way interaction between time, pitch and duration levels suggests more integration across dimensions at Post-Test II compared to Post-Test I (β =1.62, p=.009). A significant four-way interaction between time, group and levels of both acoustic features (β=-2.67, p=.007) suggests that the strength of the change in the interaction between pitch and duration cues over time differed between the groups. To follow up on a significant interaction between time, group, and pitch level, I used two regression models, one for each training group. This analysis demonstrated decrease in pitch from Pre-Test to Post-Test I for Vocabulary group (β=-.99, p<.001) and from Post-Test I to Post-Test II for Prosody group (β=-1.07, p=.002). Model 2 confirmed these results showing decrease in pitch use for both groups across time (β=-2.276, p<.001; Vocabulary$_{\text{Pre-Test vs Post-Test II}}$ β=-.639, p=.017, Prosody$_{\text{Pre-Test vs Post-Test II}}$ β=-1.758, p<.001

A


Prosody
A. Pre-Test
B. Post-Test I
C. Post-Test II
M
0.75
0.50
0.25
0.00
Duration level
Pitch level
1
2
3
4
D. Difference I
Δ AB
0.02
0.01
0.00
-0.01
-0.02
E. Difference II
Δ AC
0.050
0.025
0.000
-0.025
-0.050
F. Difference I-II
Δ BC
0.025
0.000
-0.025

B


Vocabulary
A. Pre-Test
B. Post-Test I
C. Post-Test II
M
0.75
0.50
0.25
0.00
Duration level
Pitch level
1
2
3
4
D. Difference I
Δ AB
0.00
-0.02
-0.04
E. Difference II
Δ AC
0.000
-0.025
-0.050
F. Difference I-II
Δ BC
0.02
0.00
-0.02

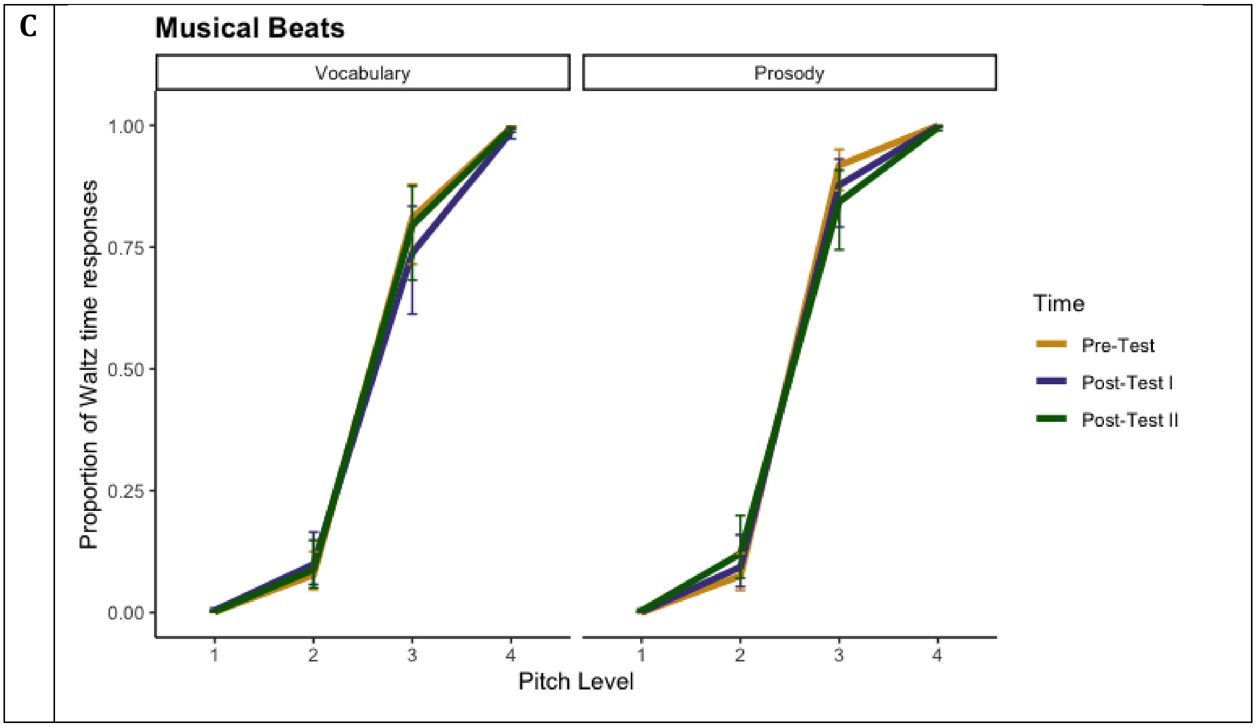


Figure 25. Musical beats categorization responses patterns. (25AB, A-F) Mean responses patterns at Pre-Test, Post-Test I and Post-Test II and the differences between the sessions plotted separately for Prosody and Vocabulary training groups. (25C) Predicted proportion of Waltz time vs March time responses for Prosody and Vocabulary training groups at Pre-Test, Post-Test I and Post-Test II. Responses averaged across participants; error bars – 95%CI.

Effects of training on prosody perception

The prosody perception trial-by-trial data was analyzed with mixed-effects logistic regression models using the glmer function from the lmer4 package (Bates et al., 2015). The dependent variables were responses on each trial (0 – incorrect, 1 – correct). I built two models for each feature – Model 1 to examine the short-term training effects and Model 2 long-term training effects. In Model 1, the testing session (Pre-Test, Post-Test I) was treatment coded (0 and 1, respectively). In Model 2, I used the contr.sdif function from the MASS package (Ripley et al., 2023) to estimate successive differences between Pre-Test (Time I) and Post-Test I (Time II) and then between Post-Test I (Time II) and Post-Test II after 6 months (Time III). Participants' unique IDs were included as a random intercept.

Results are presented in Table 26 and Figure 26. I found a significant two-way interaction between time and training group, suggesting differences in perception accuracy between groups from Pre-Test to Post-Test I (β=.309, p=.044). Post-hoc tests that involved running two separate regression models for each training group revealed that this effect was driven by decreased accuracy in performance in Vocabulary group (β=-.23, p=.003). There were no effects of training on perception of linguistic focus or lexical stress (p>.05). No long-term effects of training were found (p>.05).

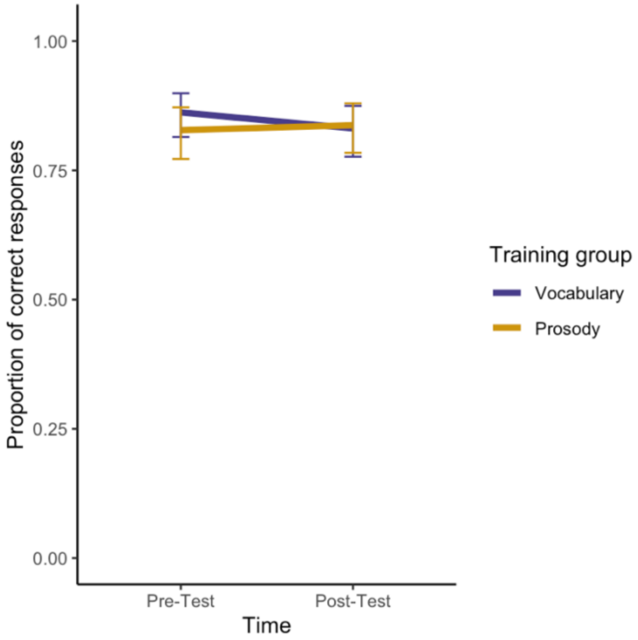


Figure 26. Predicted proportion of correct responses in Phrase Boundary perception for Vocabulary and Prosody training groups at Pre-Test and Post-Test.

Table 26. Results from mixed-effects regression analyses testing the differences between Pre-Test and Post-Test I (Model 1) and Pre-Test, Post-Test I and Post-Test II (Model 2) in prosody perception.

	Model 1				**Model 2**			
Linguistic focus	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	1.639	.194	8.433	**<.001**	1.545	.198	7.781	**<.001**
Time (I → II)	-.071	.108	-.661	.509	-.000	.126	.000	1.00
Time (II → III)	--	--	--	--	.167	.128	1.302	.193
Group (Vocabulary)	.169	.203	.830	.407	.298	.219	1.363	.173
Time (I → II) x Group	.071	.157	.456	.649	.019	.186	.103	.918
Time (II → III) x Group	--	--	--	--	-.196	.188	-1.043	.297
	Model 1				**Model 2**			
Phrase boundary	**β**	**SE**	**z**	**p**	**β**	**SE**	**z**	**p**
(Intercept)	1.833	.181	10.159	**<.001**	1.610	.179	9.009	**<.001***
Time (I → II)	-.239	.110	-2.175	**.030**	-.315	.132	-2.389	**.017***
Time (II → III)	--	--	--	--	-.115	.125	-.889	.374
Group (Vocabulary)	-.264	.147	-1.801	.072	.016	.135	.120	.904
Time (I → II) x Group	.309	.154	2.014	**.044**	.229	.185	1.239	.215
Time (II → III) x Group	--	--	--	--	.086	.179	.482	.630
	Model 1				**Model 2**			
Lexical stress	**β**	**SE**	z	**p**	**β**	**SE**	z	**p**
(Intercept)	1.602	.191	8.410	**<.001**	1.496	.181	8.274	**<.001***
Time (I → II)	.100	.110	.902	.367	.048	.126	.383	.701
Time (II → III)	--	--	--	--	.108	.128	.845	.398
Group (Vocabulary)	.227	.201	1.132	.258	.367	.200	1.937	.053
Time (I → II) x Group	-.147	.159	-.922	.357	-.162	.190	-.853	.394
Time (II → III) x Group	--	--	--	--	-.215	.188	-1.143	.253

5.4 Discussion

This study tested whether targeted perceptual training could change cue weighting strategies and whether these changes would be reflected in enhanced dimensional salience of L2-relevant dimensions and selective attention to these dimensions. In my data, I found no effects of training on dimensional salience and no dimension-specific enhancements in attention. I observed increased reliance on durational cues in the experimental group (prosody training) compared to the control group (vocabulary training), limited only to phrase boundary stimuli where duration is a primary cue. To dissociate these immediate training effects from long-term changes, I tested participants' performance 6 months after completing the study, but no long-term effects on categorizing phrase stimuli were present. These results reveal a trend towards refining cue weighting strategies following as little as three hours of training. On the other hand, participants in the Vocabulary group showed stronger reliance on pitch over time in categorizing lexical stress stimuli suggesting that in the absence of alternatives, listeners tend to solidify the use of the strategy that works well for them. Finally, I expected that changes in cue weighting strategies would lead to observable benefits in L2 prosody perception. I did not find any training improvements immediately after the training, and no lasting effects were present. Vocabulary training participants showed decreased accuracy in their phrase boundary perception, suggesting that the strategy they use might not be effective. No other effects on prosody perception were found. The results only partially aligned with my hypotheses derived from speech perception and cue weighting theories (e.g., Francis, Baldwin, & Nusbaum, 2000), suggesting that listeners can shift their relative cue weights, but perhaps more training is needed to overwrite the lifetime influences of L1 on listening strategies and achieve long-lasting effects.

Perceptual training can change cue weighting strategies

I proposed that a major source of difficulty in learning to perceive and produce L2 speech prosody is that individuals have trouble resisting default strategies inherited from their first language. I predicted that targeted training focused on enhancing salience of durational cues would result in more native-like cue weighting strategies, less biased towards pitch in categorizing stimuli where pitch is a primary cue and more towards duration where duration is a primary cue. The results from my study are mixed. On the one hand, as predicted, I observed stronger reliance on duration in categorization of phrase boundaries following targeted training in the experimental compared to control training group, but these effects were only present immediately after the training, but not after 6 months. On the other hand, participants from the control group showed a systematic increase in their reliance on pitch in categorizing stimuli for which pitch was a primary cue (i.e., lexical stress), suggesting that in the absence of any guidance their default strategies might be reinforced. These results reveal that listeners can shift their relative cue weights during L2 prosody categorization, but these changes did not apply uniformly across tasks (see results from Chapter 4 demonstrating that no single strategy is applied across all tasks). This shows that listeners can track the variability across multiple acoustic cues and dynamically adjust speech categorization to reflect statistical regularities across tasks. It is possible that I did not observe any shifts in strategies in categorizing lexical stress and linguistic focus, as duration is not a dominant cue for these contrasts. Such short exposure to artificial distributions of pitch and duration was probably insufficient to guarantee more stable refinements, as reliance on pitch, the default strategy for Mandarin speakers was already quite effective for categorizing these stimuli. It is possible that over time their strategies naturally become more biased towards pitch. Perhaps it is just easier to overcompensate in performance with a familiar strategy than learn to use a completely new one. However, a more detailed analysis testing whether participants are more likely to place a greater weight on a single primary cue or integrate across dimensions would be necessary to establish whether indeed participants become much more extreme with their strategies (e.g., decision-bound computational model, Roark & Holt, 2019). The effect of enhancement on duration use in phrase categorization could wane as while participants have completed the training, they are no longer exposed to speech with enhanced duration cues and so they might be forced to revert to strategies that worked well for them so far (i.e., more pitch-oriented).

Although I was ultimately interested in comparing participants' pre-test and post-test performance to see whether they had updated their listening strategies, individuals might have progressed through the training with different speeds and learning trajectories. Changes in strategies also do not occur suddenly after the training but happen gradually while listeners are exposed to speech sampled from acoustically manipulated stimuli space. Perceptual choices made during the training might depend not only on the current stimulus but also on the context in which the stimulus is presented or the history of participants' own choices (Urai, Braun, & Donner, 2017; Urai et al., 2019; Bosch, Fritsche, Ehinger, & de Lange, 2020). I should conduct a detailed analysis of their performance to understand how participants' behavior changed throughout the training and whether I can track the within-training shifts in strategies. Then, I will be able to answer whether learners continued using their default strategies and relying on pitch even though this dimension was unreliable or showed signs of changes throughout the training. A growth curve analysis could be applied to model multiple data points or account for individual differences (e.g., Mirman, Dixon, & Magnuson, 2008; Oleson et al., 2016) or the PsyTrack package for Python (Roy et al., 2018, 2021) – a novel method for characterizing time-varying behaviour during decision-making tasks. The latter method is based on a dynamic Bernoulli general linear model with psychophysical weights introduced at each trial to capture trial-to-trial changes. The higher the weight assigned to a particular variable, the more the participants' decision relies on that variable. I could estimate those weightings because decision-making behaviour during training might be influenced not only by the relevant variables (e.g., stimuli itself, experience with stimuli and the amount of acoustic information contained in each signal) but also by irrelevant variables (e.g., initial bias towards one of the acoustic dimensions, choice history). With this model, participants' behaviour could be quantified at a single-trial resolution to track their learning progress throughout several training days. This analysis would also allow me to gain insight into the development of their second language perceptual strategies.

Changes in cue weighting strategies do not generalize to better prosody perception

Apart from decreased performance in control group in detecting phrase boundaries, I did not find any enhancements in prosody perception following training. It could be that more exposure is needed to re-adjust the categorical boundaries between L2 linguistic classes.

One way to establish whether such changes occurred would be to look at the degree of steepness of categorization slopes (e.g., Schertz, Carbonell, & Lotto, 2019; Schertz & Clare, 2019). Such an additional analysis would tell us whether these adjustments in cue use reflect shifts of boundaries between categories. If the boundaries remained unchanged, there would be no reason to expect improvements in prosody perception.

Another important aspect to consider when designing any training is whether training particular skills was introduced at the right stage of L2 learning. For example, auditory processing training might be more useful at the initial stages of learning to facilitate the development of new phonetic categories (e.g., Saito et al., 2022). But prosodic contrasts might be conceptually more difficult and require more nuanced knowledge of L2 than learners have at the initial stages of their L2 immersion. As a consequence, if I introduced the perceptual training with prosody features too early, then participants may not have been able to benefit from it. For instance, participants who had just arrived in the UK might not have had enough exposure to native L2 speech to detect differences in prosodic contrasts or become familiar with English intonation. Perhaps such training would be better for learners with enough exposure to native L2 (e.g., longer LOR). On the other hand, if I introduced the prosody training too late, I might have run into a risk of trying to retrain fossilized strategies or re-learning incorrect or incomplete representations of L2 prosody acquired through years of learning in an L1 country. An earlier study showed that changes in cue weighting in naïve native English learners of Chinese showed rapid shifts in perceptual space during the first months of learning but plateaued as soon as after 3 months (Wiener, 2017). If that is the case, then any attempts to refine listening strategies should occur alongside the regular L2 English teaching curriculum. Unfortunately, without a longitudinal investigation that would track participants' learning trajectories from the day they started, it is impossible to discern which skills are most beneficial at which stages of L2 learning.

Also, since at the time of the second post-test some participants had already finished their studies and returned to China, it is possible they reverted to their L1 strategies after returning to their home countries (attrition of L2 after coming back to L1 country, Mickan, McQueen, Brehm, & Lemhofer, 2022). They showed a trend towards improvement while still in the UK, but if they were no longer practicing English and their daily exposure to native intonation decreased, there was not enough reinforcement from the environment to support these changes for a longer time. A more careful selection of participants that continued to reside in the UK would be helpful, although for this study, that would further reduce the number of participants. Of course, improvements in the overall accuracy of perception are only one of the possible outcomes of the training. It could be that while the accuracy is not improving as fast as I wished for, participants could have improved their processing speed. Processing speed is highly relevant for L2 learning (e.g., Hui & Godfroid, 2020) as communication usually happens in fast-paced environments where there is little time for making perceptual decisions or repetitions. I collected reaction time data across all the behavioural tasks, so a good next step would be to investigate any improvements in processing speed.

Short-term perceptual training does not lead to enhanced dimensional salience or dimension-specific attention

Finally, contrary to my expectations, I did not find training-related enhancements of dimensional salience or attention to durational cues. It could be that salience of acoustic dimensions I measure here is not representative of natural listening conditions in which participants usually make such perceptual decisions. It is also possible that three hours of training are simply not enough to overwrite a lifetime of experience with L1 distribution of salience across acoustic cues. However, because listeners cannot be automatically (exogenously) drawn to some dimensions does not mean they could not redirect their attention to these dimensions if they wanted to (endogenously). This interpretation echoes my findings from Chapter 2, where I compared L1 English and Mandarin speakers and found no tonal language-related enhancements in pitch salience, but better attention to pitch. It could be that salience is not enough to guide perception (e.g., Lengeris, 2009).

I also did not find any dimension-specific effects of training on attention. It is possible that while attention to a particular L2-relevant dimension might be needed in speech perception, the ability to adaptively switch attention between various acoustic dimensions might not be tied to specific dimensions. Acoustic dimensions within speech are also interdependent, so the performance we measure would be always relative to the other contrastive dimension. If performance across domains and dimensions is correlated, perhaps it would be more informative to consider attention to dimension as a general ability and collapse participants' performance across conditions instead of seeking out effects across multiple conditions.

Conclusions

Although the prosody training effects were limited only to phrase boundary categorization and were not stable over time, it is quite remarkable to observe that with as little as three hours of training, we could change perceptual strategies that took a lifetime to develop. It is also important to emphasize that these strategies were employed during naturalistic and complex speech perception tasks, not during the categorization of isolated syllables. These findings offer a new direction for designing more targeted language training paradigms. However, more work is needed to establish how to individualize these paradigms and efficiently select stimulus spaces appropriate for refining L1-specific language difficulties experienced by learners.

Chapter 6. General discussion

6.1 Overview of empirical findings

This thesis sought to investigate the perceptual strategies underlying second language acquisition. I conducted a series of cross-sectional and longitudinal experiments investigating how native language experience and musical training shape the salience of various acoustic dimensions and our ability to use that information in learning new languages. I focused on the two main aspects of perceptual strategies in language learning: what drives them and whether they can be changed. First, I compared how prior experience (language background, musical training, and their interaction) shapes cue weighting strategies. I did that in the context of native English speakers and native Mandarin Chinese learners of English. Then, I tested whether the weighting of different acoustic cues during speech perception reflects the direction of attention towards or salience of the most informative cues. To that end, I conducted a series of behavioural and neuroimaging studies that examined native speakers' perceptual profiles. I also examined the roles of dimension-selective attention and dimensional salience in shaping those profiles. Second, I explored whether training participants to attend to neglected perceptual dimensions could help them acquire more nativelike L2 speech perception strategies. I conducted a longitudinal study with targeted training focused on redirecting listeners' attention towards an L2-relevant acoustic cue. I examined whether this training had any consequences on participants' ability to selectively attend to acoustic dimensions and salience of those dimensions. Here, I provide an overview of the main findings.

Perceptual strategies are shaped by language background and musical training

Chapter 2 investigated the roles of language background and musical experience in shaping perceptual strategies. I asked whether there are any systematic differences in how language background and musical experience influence cue weighting strategies and whether they are linked to one's ability to selectively attend to acoustic dimensions or enhanced salience of these dimensions. Building on the assumptions of cue weighting theories (e.g., Francis et al., 2000; Francis & Nusbaum, 2002), I predicted that listeners would up-weight the dimension that is especially relevant in their L1 (i.e., pitch for conveying meaning in Mandarin Chinese) or was learnt during targeted training (i.e., pitch for conveying musical structures for trained musicians). I also expected that listeners would show superior performance in behavioural tasks requiring the use of that dimension and increased salience of that dimension reflected by its more robust cortical tracking.

My findings partly supported these predictions and revealed differences between tonal and non-tonal language speakers' behavioural performance and listening strategies. Mandarin speakers, compared to native English speakers, showed enhanced attention to and preferential use of pitch across behavioural tasks (up-weighting pitch during prosody and musical beats categorization and demonstrating superior attention to pitch). However, there was no effect of language background on neural entrainment to acoustic dimensions. I also found that although Mandarin speakers performed better on attention to pitch conditions, they did not perform worse than English speakers when asked to ignore pitch and attend to duration. One possible explanation of these results is that the mechanism underlying the increased ability to attend to acoustic dimensions might not be exogenous (i.e., driven by salience, involuntary capture of attention by stimuli) but endogenous (i.e., ability to attend to task-relevant acoustic dimensions). Comparison of cue weighting strategies between musicians and non-musicians revealed that musical training sharpens tuning to the dimension most relevant to a given categorization task. This pattern of responses suggests that the effects of musical training might depend on L1 background and highlights the distinct impact of these two types of experience on auditory perception.

Perceptual strategies need time to unfold and likely depend on multiple factors

In Chapter 3, I attempted to reconcile a long-lasting debate about the role of relative salience and dimension-selective attention in explaining the differences in cue weighting strategies. The dominant hypothesis in the field states that L1 experience inevitably leads to shifts of attention towards the most informative and reliable cues or enhances their salience (Francis & Nusbaum, 2002; Holt et al., 2018). If so, such dimensional enhancements would be reflected by assigning them stronger weights during speech perception and should explain the observed differences in cue weighting patterns. My results demonstrated that neither dimension-selective attention nor dimensional salience emerged as significant predictors of cue weighting strategies. While my study did not offer evidence to support the theory-driven predictions, we must be careful about drawing overly strong conclusions from this null result. As discussed in Chapter 3, the low reliability of my measures might be an issue here, especially since normalized cue weight is a derived factor.

I also examined whether individual differences in cue weighting strategies are stable across tasks. I found shared strategies applied to word stress and musical beats categorization, but not across other tasks. The lack of consistency in perceptual strategies across categorization tasks indicates that they are not driven by domain-general abilities such as auditory sensitivity. That also suggests that there is no uniform strategy that can be applied to any listening task, but rather there are strategies that draw on the acoustic resemblance of stimuli or relevance of cues for a given task. For example, a common characteristic of lexical stress and musical beats is their rhythmical pattern (sequences of strong vs weak notes or syllables). Listeners may optimize their strategies to capitalize on shared characteristics across tasks. Taken together, my findings suggest that the relationship between relative salience, dimension-selective attention and cue weighting might not be as straightforward as previously thought. Cue weighting strategies seem to vary depending on the specific task and other factors beyond salience and attention (e.g., processing speed, Hui & Godfroid, 2021; working memory, Francis & Nusbaum, 2009; auditory-motor integration, Kachlicka, Saito, & Tierney, 2019; rhythm perception, Slater et al., 2018) might be responsible for the observed individual differences.

Dimension-selective attention is linked to better speech perception and grammatical knowledge

The goal of Chapter 4 was to test whether and to what extent individual differences in L2 learning experience, auditory abilities and dimensional salience can explain the variability in L2 learning outcomes across several aspects of language competence: perception and production of L2 speech prosody and lexical and grammatical knowledge. My findings provide support for cue weighting theories implicating the role of attention in L2 speech acquisition (e.g., Francis & Nusbaum, 2002; Holt et al., 2018). I show that attention to acoustic dimensions emerged as a significant predictor of not only vowel and prosody perception, but also grammar knowledge, extending its role from spoken language to syntax. Furthermore, my research contributes to the ongoing discussion about the neurocognitive model of language aptitude (Turker, Seither-Preisler, & Reiterer, 2021; Turker & Reiterer, 2021). My results lend support to the idea that no single factor can explain variance in L2 learning success as L2 learning is a complex process that encompasses a combination of innate characteristics, experiential factors, cognitive faculties, neural mechanisms and even genetic predispositions (Turker et al., 2021). Integrating attention into this debate is one step towards a more integrative model of L2 proficiency that combines various facets of that process and acknowledges that different factors not only might play a role in mastering different aspects of language but might also be relevant at different stages of L2 learning and have differential effects depending on the individual.

Perceptual strategies are susceptible to training

Finally, in Chapter 5, I set to establish whether it is possible to shift perceptual strategies towards an L2-relevant dimension and boost L2 learning with targeted training emphasizing the role of that dimension in speech perception. I developed a novel training paradigm that aimed to boost salience of durational cues, and by doing so, shift listeners' attention towards duration by suppressing competing pitch information. Prosody training entailed listening to speech samples modified in fundamental frequency (F0 perceived as voice pitch) and duration and categorizing them as belonging to one of two linguistic contrasts spanning linguistic focus, phrase boundary and lexical stress. Acoustic manipulations included removing F0 entirely by creating whispered-like speech and then gradually introducing F0, which was present but never task-relevant, and decreasing the size of the duration cues, thus increasing the difficulty of the categorization task over the course of the training. Vocabulary training was matched in length and intensity to prosody training and involved non-auditory English vocabulary learning tasks.

I demonstrated that it is possible to shift perceptual strategies during L2 speech prosody perception following short-term targeted training. After as little as three hours of training, I observed stronger reliance on duration in categorization of phrase boundaries. Even though these effects were only present immediately after the training, but not after 6 months, participants were able to generalize their training to new exemplars they had not heard before. My findings suggest that listeners can track the variability across multiple acoustic cues and dynamically adjust speech categorization to reflect statistical regularities across tasks. Nevertheless, pitch reliance is the default strategy for Mandarin speakers, so overturning a lifetime experience to establish more stable refinements in using durational cues might necessitate more training. Overall, these findings provide evidence of perceptual learning at the suprasegmental level, extending the existing literature showing similar effects on segmentals (e.g., Kondaurova & Francis, 2010; Zhang, Wu, & Holt, 2021).

6.2 Limitations and future research

My empirical investigation into the perceptual strategies underlying second language acquisition has provided valuable insights into the roles of language background, musical experience, attentional mechanisms, and training in shaping individuals' cue weighting strategies during speech perception. While the findings have shed light on certain aspects of L2 learning, it is essential to acknowledge the limitations of the present research and identify potential directions for future studies. This section reviews the limitations and proposes solutions for advancing our understanding of perceptual strategies in second language acquisition.

Neural measure of dimensional salience

Existing attention-to-dimension theories of speech perception suggest that dimensions that are important for conveying structure in an individual's L1 become particularly salient or likely to capture attention (Francis & Nusbaum, 2002; Gordon, Eberhardt, & Rueckl, 1993; Holt et al., 2018*)*. According to these models, this acquired salience of the language-specific dimensions might lead to upweighting of that dimension during perception and drive increased attentional gain to that dimension. I tested this hypothesis in the context of Mandarin Chinese native speakers who, unlike English speakers, make use of L1-relevant pitch variations to convey word meaning in their native language. In Chapter 2, I compared dimensional salience between Mandarin and English speakers, and I predicted that Mandarin speakers would show enhanced pitch salience. Then in Chapter 5, I attempted to boost Mandarin speakers' L2 learning by enhancing the salience of the competing dimension – duration with targeted training. However, in both cases, I found no evidence to support my hypotheses – there were no differences in dimensional salience between Mandarin and English speakers and no changes in salience after the training.

I used the EEG frequency tagging paradigm and computed inter-trial phase coherence (ITPC) as a measure of dimensional salience. When the listeners listen passively to streams of sounds changing in two acoustic characteristics at two different rates, more salient characteristics are meant to result in stronger ITPC at a given frequency of change. For example, an earlier study showed enhanced tracking for 2-semitones compared to 1-semitone pitch step sizes, whereas spectral peak tracking was unaffected by the step size (Symons et al., 2021). Earlier studies also showed that increases in pitch salience (defined as the increase in temporal regularity of iterated ripple noise) result in increased signal amplitude as measured with magnetoencephalography (Krumbholtz et al., 2003) or increased firing rates of auditory cortical neurons during electrocorticography (Schonwiesner & Zatorre, 2008).

One problem with using such an approach for measuring language-relevant salience is the stimuli I selected for this study. Namely, they were verbal and non-verbal sounds, but the verbal stimulus was a single isolated vowel repeated multiple times. A single vowel does not resemble natural acoustic variations of naturalistic speech in any way. If the salience of acoustic dimensions is learned as a part of L1 exposure and tied to its linguistic context, it is reasonable to expect that context is necessary to elicit speech-relevant bottom-up salience. Another issue is that I used the same base stimuli to create sequences for frequency tagging as those used for the dimension-selective attention task. However, the stimuli properties were carefully selected for attention tasks to be balanced for salience. If the salience of acoustic dimension is estimated relative to its background or competing source of information, perhaps I disturbed the natural way these acoustic dimensions would capture listeners' attention by manipulating them. Consequently, these artificial changes to stimuli properties were not salient enough to detect and be reflected in neural tracking, especially if they lacked linguistic relevance to participants. One way to test these assumptions would be to conduct a follow-up study using an experimental paradigm similar to the one reported in this study but with linguistically meaningful stimuli (e.g., one-syllable words). It would also be wise to test the tracking of acoustic dimensions in listeners' L1. If the assumed pitch salience amongst Mandarin speakers cannot be reliably tracked with such a paradigm while they are listening to native speech, perhaps this is not the right way of measuring dimensional salience. Alternatively, it might be that the L1 experience does not lead to enhanced salience of acoustic dimensions.

Furthermore, salience is a bottom-up mechanism driven by distinctive or task-relevant stimulus characteristics, so it seems plausible that it would be tied to the fidelity of neural encoding of those features. Previous studies looking into the relationships between subcortical, cortical, and behavioural measures of pitch salience indicated that information related to pitch salience might emerge early along the auditory pathway and likely originates from the pre-attentive, sensory-level processing (Krishnan, Bidelman, & Gandour, 2010; Krishnan et al., 2012). A study using iterated rippled noise showed that pitch discrimination improvements with increasing salience were reflected by increased precision of neural pitch encoding (Krishnan et al., 2010). Research also found that behavioural responses were highly correlated with a greater magnitude of cortical responses and shorter response latencies (Krishnan et al., 2012), a relationship I did not observe in this study. As suggested above, it could be that the differences in pitch levels were not salient enough to elicit such a response. However, since tonal language speakers are known to benefit from enhanced pitch processing (e.g., more precise pitch discrimination, Giuliano et al., 2011; better melody discrimination, Liu, Hilton, Bergelson, & Mehr, 2023), it is possible they would show enhanced pitch encoding (e.g., using frequency-following response; Krishnan, Gandour, & Bidelman, 2010). This could also be the case for musicians – it was also shown that musical experience could enhance pitch encoding (Wong et al., 2007).

Encoding of pitch information at the early levels of the auditory pathway could be limited to individual harmonics that form a general representation of pitch in the auditory cortex (Plack, Barker, & Hall, 2014). If so, there is plenty of ways in which cortical pitch representations could be mediated by additional sensory and extrasensory influences, including top-down attention (for discussion about attention effects at different stages of auditory processing see Sussman, 2017; Price & Moncrieff, 2021). A further study that tracks neural responses at both subcortical and cortical levels could reveal whether tone language experience or musical training can boost the processing of early pitch encoding without affecting pitch salience. Another way to distinguish purely acoustic processing from higher-level perceptual or linguistic representations would be to track the encoding of acoustic, phonetic and linguistic pitch-related features during naturalistic speech perception. More ecologically valid approaches, such as techniques using mTRF (e.g., Teoh et al., 2019; Di Liberto et al., 2020; Brodbeck & Simon, 2022), would allow us to measure the strength of encoding of various features at different time lags, possibly helping to disentangle individual contributions of various components of pitch to speech perception. Additional research could also compare the encoding of these features in native vs non-native speech to pinpoint the differential effects of various L1 backgrounds on the processing of acoustic information.

It could also be that experience with pitch in speech perception is reflected in increased reliability of that dimension. Previous studies on multimodal integration suggested modality reliability as a potential mediating factor of functional connectivity, and results conformed with that assumption showing cue weighting shifts towards more the reliable modality in syllable perception (Beauchamp et al., 2010; Nath & Beauchamp, 2011). Jasmin and colleagues (Jasmin, Dick, Stewart, & Tierney, 2020) investigated cue reliability in the context of patients diagnosed with congenital amusia – a condition that primarily affects pitch memory (Tillman et al., 2016) rather than low-level pitch processing (as evidenced by, e.g., similar activation of pitch-responsive areas, Norman-Heignere et al., 2016, comparable pitch strength measured with FFR, Liu, Maggu, Lau, & Wong, 2015). The results showed that the reliability of perceptual dimensions is linked to functional connectivity between frontal and perceptual regions and suggests compensatory mechanisms (Jasmin et al., 2020). It is plausible that tonal language speakers, who throughout their lifetime experience listening to pitch contours marking the meaning of every word, would acquire higher pitch reliability and show increased connectivity patterns between these regions when pitch cues are present.

Factors beyond salience

I also found that neither dimension-selective attention nor dimensional salience predicts cue weighting strategies. This null result could be attributed to the low reliability of the measures I used. Regardless of whether dimensional salience and attention are involved, several other factors could play a role in shaping cue weighting strategies. Previous studies showed that people are distracted by the increased attentional load (completing arithmetical tasks, Gordon et al., 1993) or the presence of competing talkers (Symons, Holt, & Tierney, 2023) and as a consequence, might adaptively shift attention in response to demanding listening conditions. These results point to the possibility that cognitive abilities that reduce distractibility and enhance attentional focus (e.g., processing speed, Hui & Godfroid, 2021; working memory, Francis & Nusbaum, 2009; auditory-motor integration, Kachlicka, Saito, & Tierney, 2019; rhythm perception, Slater et al., 2018) could be relevant in shaping perceptual strategies. For example, increasing working memory workload was shown to impede speech recognition in terms of reaction time and accuracy (Francis & Nusbaum, 2009), so it is plausible to assume that better working memory could have some beneficial effects.

Detailed examination of musical training effects conducted in Chapter 2 emphasized the role of training in shaping listening strategies. However, musicians are only one group of listeners that can have better than average auditory skills (e.g., Micheyl et al., 2006; Zaltz, Globerson, & Amir, 2017). There are also people who are naturally very musical or have good auditory skills. In an EEG study, people who did not obtain any musical training, but had better listening skills (called by authors "musical sleepers") presented with stronger categorical processing around the time frame of P2 (~180 ms) in the right auditory cortex and were behaviorally faster and more accurate compared to poorer listeners (Mankel, Barber, & Bidelman, 2020). This suggests that not only training, but inherent listening abilities or musicality can contribute to better and more efficient sound-to-meaning mapping. Such an ability would be of course very useful for aspiring musicians since learning musical notation heavily relies on sound-to-meaning mapping.

Generalizability of training effects

Although I observed increased duration use in categorizing phrase boundaries following targeted training, these effects were not maintained over time. Furthermore, I found that these durational enhancements in cue weighting did not translate to better phrase boundary perception. Below I discuss several factors that might be responsible.

The balance between ecological validity and acoustic control

In the training design, I opted for naturalistic stimuli to guarantee a maximally engaging and ecologically valid paradigm. I used natural recordings encompassing natural durational variations across linguistic contrasts, and I also made sure to include a variety of male and female voices with distinct pitch ranges and of different ages. Although all that contributed to the high quality and variability of speech samples in the training, it also introduced a lot of uncontrolled variability across other acoustic dimensions. As mentioned earlier, combinations of cues in cue weighting cannot be interpreted independently, as they covary consistently with one another (i.e., dimensional integrality; Francis, Kaganovich, & Driscoll-Huber, 2008). In other words, changes implemented along one dimension inevitably lead to changes across other dimensions. I attempted to control those potential confounds during the morphing procedure by keeping parameters besides duration constant across all training samples. But even if the acoustic cues of interest are controlled by morphing procedures, it could be that these changes disturb the natural order of weights during speech perception. These artificial changes could likely introduce unwanted hints from secondary dimensions and make the target acoustic space difficult to describe. One simple solution to that problem would be to conduct an acoustic analysis across all speech samples included in the training to map how they are represented by the acoustic dimensions of interest. I could then establish the presence of boundaries between the categories and their exact descriptors (for a similar approach to phonetic categories, see e.g., Iverson & Kuhl, 1995). Multidimensional scaling mapping would also allow me to tag those samples that extend beyond the average space occupied by a given category.

Alternatively, using synthesized speech could have guaranteed complete control over speech acoustics. That would also alleviate the issue of coarticulation and context on cue weighting (Schertz et al., 2019), as information external to the relevant sounds has also been shown to influence the perception of linguistic contrasts (Repp, 1982).

Since phrase boundary stimuli were the only stimuli in the training where duration was a primary cue (compared to lexical stress and linguistic focus where pitch is a primary cue), a better selection of stimuli could have been made. Presumably, additional trials that included examples spanning contrasts in which duration bears more relevance than other cues would have a better effect on training perceptual strategies. One way to achieve that would be to focus on the phrase boundary examples and increase the number of trials. But one could also include a variety of other tokens that can best be resolved by relying mainly on duration. For example, in English, duration is a primary cue for distinguishing between long and short vowels, voiced and voiceless fricatives, or phrase-final and non-final syllables (Klatt, 1976) that could be incorporated into the training materials.

Another crucial aspect related to acoustic control is the fact that I kept the natural pauses between phrase boundaries in the phrase boundary stimulus set. The intention here was to keep the samples as natural as possible. I acknowledge this might not have been the optimal choice because although pause length could be considered a durational cue, it might also be processed separately (Scott, 1982). The pause occurring before a phrase boundary is indeed perceived as a distinct cue and may be assigned different weights in perception. For instance, while pauses may have limited relevance in German (Wellman et al., 2012), Mandarin speakers tend to attribute greater weights to pauses than to durational cues, at least when processing their native language (Yang, Shen, Li, & Yang, 2014). Therefore, the presence of pauses could act as a significant confound to my results potentially attenuating the effects of prosody training. In retrospect, I believe that removing pauses from the stimuli would have been a better choice, and I recommend that future studies explore the effects of training without pauses – in fact, such a follow-up study is already underway.

Extending training effects on perceptual strategies

An ultimate goal of L2 learning for most learners is to communicate effectively with other people and communication by definition includes some form of interaction with another person and exchange of information. Previous research emphasized the role and potential benefits of social context in language learning (e.g., Li & Jeong, 2020; Jeong et al., 2021). However, my training paradigm focused entirely on unidirectional categorization tasks that lacked any elements of interaction. While designing training that could be applied at a large scale or remotely and would also incorporate actual face-to-face interaction is an impossible task, including some interactive elements to daily exercises could boost participants' motivation and potentially their learning outcomes. For example, embedding the stimuli within short stories where participants need to answer correctly to uncover another piece of missing information to understand the whole story would make it more engaging. A simulated conversation with chat-like features that uses response options and prerecorded messages would also serve that purpose. To provide a more holistic learning experience, I could also include tasks that prompt participants to speak. For example, it would be easy to incorporate delayed or immediate recall tasks to practice production. However, research suggested that while training perception might have beneficial effects on production, production during training disrupted perceptual learning (Baese-Berk, 2019).

Some other choices I made at the design stage could also influence the effects of training. For example, no explicit instruction as to the purpose of the training was included, and participants had to implicitly learn to rely on durational cues while making categorization judgements. Since the aim of the training was somewhat concealed (i.e., I did not explicitly inform participants that they should be paying attention to duration), the exercises might have seemed to be completely arbitrary. Anecdotal evidence from discussions with participants and short feedback they gave in the post-training questionnaire suggest that more emphasis on the purpose of the training would make the goals clearer to the learners and could benefit learning. It would be straightforward to explicitly instruct participants to pay attention to the specific acoustic dimension relevant to a given language. While such perceptually focused instruction may aid L2 speech learning (e.g., Yang & Sundara, 2019), this kind of instruction is usually not available in naturalistic learning conditions neither while learning a first language nor in most cases of second language acquisition (Francis, Baldwin, & Nusbaum, 2000; Francis & Nusbaum, 2002). It is possible that due to the differences in cues relevance across languages, learners would not be able to follow such an instruction. This is because they might not show sensitivity to L2-relevant dimensions, or the relative salience of other dimensions is much stronger and attracts their attention instead. To compensate for any disadvantages, the training would require lengthy explanations of speech acoustics and examples of how these acoustic features are represented in natural speech. Another downside of explicit instruction is that it might only work in certain individuals who have the meta-cognitive skills to intentionally take on new explicit strategies. Instead, I could include a familiarization session that presents several trials with untransformed stimuli to introduce listeners to the target range of acoustic features they will be dealing with. Previous studies showed that as little as a single trial could have facilitatory effects on subsequent training (e.g., a single informative presentation was sufficient to enable learning to detect oddly oriented visual elements under difficult conditions; Ahissar, 1999).

Finally, what appears to be crucial for stable learning effects is the appropriate length of training. There is limited research on L2 prosody acquisition, yet alone on prosodic cue weighting. It should be a priority to establish how much training is needed to permanently shift perceptual strategies during prosody perception via targeted training and also find a way to provide long-term learning support to retain these strategies over time. It is important because some learners learn L2 in the L1 context and do not have enough exposure to L2 speech or return to their home countries after a period of immersion, significantly limiting L2 input that could reinforce the strategies they have learned while living abroad. I argued that since, at the time of the second post-test, some participants had already finished their studies and returned to China, it is possible they reverted to their L1 strategies after returning to their home countries (attrition of L2 after coming back to L1 country, Mickan, McQueen, Brehm, & Lemhofer, 2022). My participants showed a trend towards improvement while still in the UK, but if they were no longer practicing English and their daily exposure to native intonation decreased, there was not enough reinforcement from the environment to support these changes for a longer time. It is therefore of paramount importance to design paradigms not only for training but also for long-term retention of newly acquired skills.

Individual differences

My research demonstrated that the differences in cue weighting are evident at the level of comparing speakers of different L1s. In Chapter 2, despite the emerging trend for Mandarin speakers being more biased towards pitch than English speakers, I also noted huge individual variability within both groups in what features listeners relied on during categorization tasks. These differences contribute to differential weighting patterns in speech perception and can have consequences for learning a new language and adapting listening strategies.

Individual differences in perceptual strategies

Obscuring individual differences might be a potential limitation of the relative measures employed in Chapter 3. I also did not conduct more detailed analyses of cue weighting patterns in other chapters. This is problematic because choosing responses at the two endpoints of the scale or corners of the stimuli space (i.e., categorical pattern of responses) and sampling more uniformly from the stimulus space (i.e., more gradient pattern of responses indicative of less clear boundaries between categories) would represent dramatically different listening strategies, yet they might result in the same normalized cue weights since the individual contributions of each dimension and choices across the stimuli space are somewhat blended together (Schertz & Clare, 2019). In a similar fashion, when averaging responses across the whole space, we might be diluting the differences present in the corners of the space where two cues conflict. If a listener has a strong bias towards pitch, their responses might be less accurate when pitch contribution to the linguistic contrast of interest becomes ambiguous (e.g., the amount of pitch is reduced compared to another dimension). It would be informative to look at patterns of responses where the two acoustic cues of interest conflict to establish whether we can observe such patterns of gradience vs categoricality. To achieve that we could look at the slopes of categorical responses across tasks, an approach successfully applied to investigating categorical boundaries between speech categories (e.g., Schertz & Clare, 2019).

When looking at the performance of L2 learners in more detail, L2 proficiency (as measured by a composite test of listening, vocabulary and reading comprehension) was shown to be an important factor in explaining differences in F0 vs VOT weighting patterns of English learners of Korean (Kong & Edwards, 2015). Better L2 proficiency in Korean was linked to increased sensitivity to F0 while categorizing L2 voicing contrasts compared to participants' L1 weighting, a pattern that suggests an existing interaction between L1 and L2 listening strategies at the individual level. These findings were replicated with Korean learners of English (Kong & Kang, 2022). Within this group, more proficient speakers of L2 English were shown to be less sensitive to F0 and had more categorical response patterns (i.e., target categories were more clearly separated from one another). These patterns extended to production, demonstrating that highly proficient L2 speakers use the primary cue, VOT, more effectively than less proficient learners (Kong & Yoon, 2013). Overall, these results suggest that more proficient L2 learners could be more adept in relocating their attention to relevant acoustic cues or inhibiting distractions coming from irrelevant cues. It is not yet known whether the observed gradience patterns persist over time, and so a detailed analysis of individual response patterns before and after the training would enrich insights from the study presented in Chapter 5.

Listeners might also differ in their default strategies – they might place greater weight on a single primary acoustic cue or systematically integrate across dimensions. A recent study indicated that trained musicians are more likely to rely on a single dimension during prosody categorization, while non-musicians were shown to integrate across dimensions (Symons & Tierney, 2023). Authors argued that this might be due to general auditory enhancements driven by their musical expertise. It remains unexplored whether the specialization in pitch perception arising from tonal language experience would have a similar effect. Overall, the L2 speech perception patterns tend to mirror listeners' L1 strategies. If that is the case, then we should observe more reliance on a single cue, pitch, amongst tonal language speakers. Research also demonstrated that in both L1 and L2 perception, listeners' gradient responses were linked to greater reliance on language-specific redundant cues (i.e., F0 in L2 English and VOT in L1 Korean; Kong & Kang, 2022). Further research should aim to reconcile whether these patterns are indeed consistent across language backgrounds.

Consequences of individual differences on learning trajectories

As discussed in Chapter 5, conducted analyses were insufficient to offer insights into how perceptual strategies might emerge over time and what consequences the initial patterns of cue weighting have on participants' learning trajectories. Certainly, these differential patterns of cue weighting and default listening strategies might predetermine what strategies are employed during the training. First of all, due to the inherent differences in cue weighting patterns and preferred strategies, learners do not start learning a language from the same place. If, as suggested before, there is some overlap in what strategies people apply in their L1 and while learning subsequent languages, then their default strategy that is effective for their L1 might put them at a disadvantage in learning a new language. For example, more gradient responses were linked to greater reliance on secondary cues and worse L2 proficiency (e.g., Kong & Kang, 2022). As a consequence, learners might significantly differ in their training needs. While people with more categorical responses and acquired flexibility to switch between the relevant acoustic cues might only need some fine-tuning of their strategies (e.g., refining the boundaries between the categories), others might need much more work. They might need to learn how to attend to an entirely novel acoustic property, separate new L2 linguistic contrast from similarly sounding L1 categories or learn how to use acoustic information more flexibly depending on the task at hand. If learners' training needs are not aligned with training goals, they might not succeed. Those who might need additional tuning might reach a plateau as no new or challenging material is presented to them. In contrast, those who need to develop their L2 strategies from scratch might struggle with basic exercises. That, in turn, could lead to short-term training effects limited to training material that do not generalize to other tasks or stimuli. Therefore, it is crucial to complement the analyses in Chapter 5 with a systematic analysis of individual patterns and trial-by-trial responses in search of systematic differences across people. This potentially very informative extension of my work should test if any clusters of similar behaviours emerge from the data. Such information could offer guidance about how to better tailor future training paradigms to exact learners' needs.

Towards theory building

While the research reported in this thesis brings novel insights, we must not forget that the observations made when gathering data for our experiments should be interpreted with appropriate theories in mind (Press, Yon, & Hayes, 2022). As emphasized throughout the thesis, my work fits well within the discussion of speech perception and cue weighting theories. My findings expand upon the existing ideas by showing that the roles of dimension-selective attention and dimensional salience are not as straightforward as these theories predicted. But to better understand the observed patterns and why certain predictions were not supported, it is necessary to step back, thoroughly consider and integrate what we already know.

In Chapter 1, I reviewed the most prominent theories in the field, namely the Perceptual Assimilation Model (PAM, Best, 1993, 1994; its extension PAM-L2, Best & Tyler, 2007), which incorporates both contrastive phonological and noncontrastive phonetic influences from L1 addressed earlier by the Speech Learning Model (SLM; Flege, 1986, 1995) and Native Language Magnet Model (NLM; Kuhl, 1991; Kuhl et al., 1992; Iverson & Kuhl, 1996). I also talked about the perceptual interference account (Iverson et al., 2003) that extends the concepts of perceptual nonuniformity proposed by Kuhl's Perceptual Magnet (NLM; 1991) and offers a more detailed explanation of how perceptual warping of acoustic space affects speech perception. Finally, I introduced the attention-to-dimension models that assume a spatial representation of perceptual space defined by a range of acoustic dimensions, where each of these dimensions represents a feature along which linguistic categorization can be made. The dimensional warping of that space is operationalized here in terms of attentional operations and formalized as weights or multipliers that stretch or shrink the dimensions depending on attentional focus (Francis & Nusbaum, 2002). For example, focusing attention on a given dimension increases its weight and stretches it to make the differences along that dimension more easily perceivable. While all these approaches can explain some aspects of the re-structurization of phonological space observed in language learning, they do not necessarily agree on how this re-organization would occur. Take, for example, the dimensional stretching within-category space and the shrinkage of perceptual space mentioned above. This idea is somewhat similar to what Kuhl and collaborators (Kuhl, 1991; Kuhl et al., 1992; Iverson & Kuhl, 1996) proposed in their Native Language Magnet Model, but the warping described there was localized in nature (i.e., describing shrinkage nearby phonemic category centres). Yet there is no clear consensus on how predictions made by both theories could be tested or invalidated. Second Language Linguistic Perception theory (L2LP, Escudero, 2009) and its revision (Van Leussen & Escudero, 2015) also seem to be relevant here as they extend what the other three models proposed by trying to understand the whole speech perception learning process (from naïve L2 listeners to advanced learners). This model also attempts to separate the effects of pre-lexical perception and lexical recognition in L2.

Furthermore, these are only the leading ideas in speech perception research, and at least a dozen other views should be scrutinized. For example, there is the Universal Perceptual Model (Georgiou, 2021ab), which is a direct extension of PAM and SLM discussed above, the Feature Hypothesis (McAllister, Flege, & Piske, 2002) echoing assumptions of the SLM model, or Bohn's Desensitization Hypothesis (Bohn, 1995) and many others (e.g., Full Transfer & Full Access model, Schwartz & Sprouse, 1996; Optimal Perception Hypothesis, Escudero, 2009; attention and associative learning and analyzer theory, Sutherland & Mackintosh, 1971). Additional comparisons could be made with existing computational models (e.g., TRACE model of speech perception based on principles of interactive activation, McClelland & Elman, 1986; supervised error-driven learning and reinforcement learning; Harmon, Idemaru, & Kapatinski, 2019; automatic selective perception routines models, Strange, 2011; second language linguistic perception model [L2LP], van Leussen & Escudero, 2015). These lists are obviously not exhaustive.

Of course, a thorough overview and testing of existing theories and models of speech perception is way beyond the scope of a single thesis. Nevertheless, to further research on this topic, a solid systematic review of existing theories is necessary to provide a sound background for answering any lingering questions. A comprehensive research plan should be developed to review, summarize and compare existing theories and models of speech perception and production to establish to what degree they agree or disagree with their predictions and what these predictions can or cannot explain, and finally, design studies that would test these predictions and validity of these theories.

6.3 Summary and conclusion

Regardless of the range of limitations I discussed above, this research provides valuable insights into perceptual strategies underlying L2 acquisition. The presented studies led to several important conclusions. Firstly, my results highlighted that language background and musical training have discernible influences on perceptual strategies. Secondly, the role of salience and dimension-selective attention in this process was not supported and warrants further investigation. Thirdly, I show that the role of dimension-selective attention extends from speech perception to morphosyntax. Lastly, the study demonstrates that with as little as three hours of targeted perceptual training, it is possible to adjust L2 cue weighting strategies. I show these effects for the first time in the context of suprasegmental perception. I also emphasized the importance of considering individual differences and attentional mechanisms in updating theoretical accounts related to L2 acquisition. Overall, this research underscores the complexity of the perceptual aspects involved in language learning and advocates for more nuanced and rigorous approaches that could acknowledge the diverse ways individuals process and acquire new languages.

Theoretical contributions

My findings provide novel insights into the long-debated role of dimension-selective attention and dimensional salience in shaping cue weighting strategies (e.g., Francis & Nusbaum, 2002). While I did not provide evidence supporting the hypothesis that attention and salience contribute to individual differences in cue weighting, I considered several issues that might have contributed to such a null result. This question remains open for discussion as the methodological issues above prevent us from drawing overly pessimistic conclusions. Most importantly, I demonstrated that native Mandarin Chinese learners of English did not show enhanced pitch salience but did show an enhanced ability to attend to pitch. This finding suggests that the mechanism underlying their increased ability to attend to pitch might not be exogenous (i.e., salience, involuntary capture of attention by stimuli) but rather endogenous (i.e., driven by their ability to attend to pitch if they want, but only when it is task-relevant). Further research is needed to test this proposal in more detail.

Another important contribution of this thesis is the demonstration that dimension-selective attention supports not only speech prosody perception but also extends to the acquisition of implicit grammatical knowledge. This is in agreement with proposals that perception of prosodic cues can enhance rule learning by facilitating segmentation (de Diego-Balaguer, Rodriguez-Fornells, & Bachoud-Levi, 2015) or that the production of grammatical functions can be constrained by prosodic competence (e.g., Prosodic Licensing Hypothesis, Demuth, 2007; Prosodic Transfer Hypothesis, Goad & White, 2019).

Practical contributions

The major practical contribution of this project is the development of the L2 training paradigm fine-tuned to specific difficulties in speech perception experienced by Mandarin Chinese learners (i.e., difficulties with disengaging from pitch information). These training materials could be potentially adapted to target specific acoustic contrasts people have trouble with. For example, I could introduce formant variations to facilitate learning of /r/ vs /l/ contrasts among Japanese speakers. Such a tool would allow struggling learners to improve their listening skills and boost L2 learning. There have also been claims that children with language learning difficulties (e.g., dyslexia, Leong & Goswami, 2014; SLI, Cumming, Wilson, & Goswami, 2015) have specific problems with perception of rhythmic (durational) patterns in speech, and these training materials could potentially be adapted to remediate these difficulties as well.

Although vocabulary training was only developed as a control training group, it has undeniable potential. These materials could support daily L2 learning curricula at schools and during language courses. Words included in the training could be easily organized into thematic groupings so learners could practice specific vocabulary sets relevant to a given lesson topic. Also, difficulty levels could be arranged into exercise clusters more tailored for proficiency levels. For example, easy words from earlier levels would be appropriate for beginners. To make the training more general, a straightforward extension would be to add audio recordings of all words so that learners can also practice their pronunciation. Applications like this are gaining increasing popularity these days, so tailoring training material to support regular teaching would be beneficial not only for researchers (i.e., they could track progress along the class learning curriculum) but also for teachers and students as well as it would give them more tools to learn and practice.

Conclusion

This thesis demonstrated that learning a new language requires listeners to acquire appropriate perceptual strategies that facilitate using acoustic cues most relevant to a given task. These strategies rely on cues' reliability and informativeness, the accuracy of their encoding, and contextual influences. The development of such strategies is shaped by our L1, but it can be modulated by L2 experience and targeted training. However, the mechanisms underlying changes during such training and its short-term and long-term effects on cue weighting in speech perception are largely unexplored. With this research, I provided some answers and contributed to a better understanding of L2 speech learning mechanisms and more generally to unravelling the complexities of relationships between dimension-selective attention, dimensional salience and speech perception. These results have the potential to inform the development of supportive treatments for language learning deficiencies or boosting L2 learning by targeting specific auditory challenges experienced by learners. This work could also lead to more accurate language aptitude batteries for determining whether individuals are likely to be successful language learners and which learners will need extra assistance. Furthermore, an insight into the relative cue weighting in detecting phrase boundaries, linguistic focus and lexical stress might offer an additional dimension of information crucial for solving the problem of automated prosody detection from speech recordings.

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Ahissar, M. (1999). Perceptual learning. *Current Directions in Psychological Science*, *8*(4), 124–128. https://doi.org/10.1111/1467-8721.00029

Alain, C. (2007). Breaking the wave: Effects of attention and learning on concurrent sound perception. *Hearing Research*, *229*(1–2), 225–236. https://doi.org/10.1016/j.heares.2007.01.011

Alías, F., Socoró, J., & Sevillano, X. (2016). A Review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, *6*(5), 143. https://doi.org/10.3390/app6050143

Amari, S., Cichocki, A., & Yang, H. H. (1995). A new learning algorithm for blind signal separation. In: M. C. Mozer, M. I. Jordan, & T. Petsche (Eds). *Advances in Neural Information Processing Systems*, 757–763.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arnfield, S., Roach, P., Setter, J., Greasley, P., & Horton, D. (1995). Emotional stress and speech tempo variation. *Proceedings of the ESCA/NATO Workshop on Speech Under Stress*, 13–15.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, *19*(4), 537–552. https://doi.org/10.1017/S0142716400010353

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor - International Journal of Humor Research*, *16*(2), 243–260. https://doi.org/10.1515/humr.2003.012

Baek, H. (2022). Prosodic cue weighting in the processing of ambiguous sentences by native English listeners and Korean listeners of English. *The Journal of the Acoustical Society of America*, *151*(1), 158–167. https://doi.org/10.1121/10.0009171

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*(4), 981–1005. https://doi.org/10.3758/s13414-019-01725-4

Baese-Berk, M. M., Chandrasekaran, B., & Roark, C. L. (2022). The nature of non-native speech sound representations. *The Journal of the Acoustical Society of America*, *152*(5), 3025–3034. https://doi.org/10.1121/10.0015230

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., Egorov, D., Mailhammer, R., Grant, A., & Holman, E. W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology, 13*(1), 169–181. https://doi.org/10.1515/LITY.2009.009

Bartels, C. (1999). *The intonation of English statements and questions: A compositional interpretation.* Routledge.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, *30*(6), 644–663. https://doi.org/10.1016/0749-596X(91)90030-N

Beauchamp, M. S., Pasalar, S., & Ro, T. (2010). Neural substrates of reliability-weighted visual-tactile multisensory integration. *Frontiers in Systems Neuroscience, 4*, 25. https://doi.org/10.3389/fnsys.2010.00025

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bent, T., Bradlow, A. R., & Wright, B. A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 97–103. https://doi.org/10.1037/0096-1523.32.1.97

Berti, S., Roeber, U., & Schroeger, E. (2004). Bottom-up influences on working memory: Behavioral and electrophysiological distraction varies with distractor strength. *Experimental Psychology*, *51*(4), 249–257. https://doi.org/10.1027/1618-3169.51.4.249

Best, C. T. (1993). Learning to perceive the sound pattern of English. *Advances in Infancy Research*, *9*, 31–80.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In: J. C. Goodman & H. C. Nusbaum (Eds). *The Development of Speech Perception*, 167–224.

Best, C. T. (2019). The diversity of tone languages and the roles of pitch variation in non-tone languages: Considerations for tone perception research. *Frontiers in Psychology*, *10*, 364. https://doi.org/10.3389/fpsyg.2019.00364

Best, C. C., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*(2–3), 183–216. https://doi.org/10.1177/00238309030460020701

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementaries. In: O. S. Bohn, & M. Munro (Eds.), *Second-language speech learning: The role of language experience in speech perception and production: A festschrift in honor of James E. Flege* (pp. 13–34). Amsterdam: Benjamins.

Bharadwaj, H. M., Lee, A. K. C., & Shinn-Cunningham, B. G. (2014). Measuring auditory selective attention using frequency tagging. *Frontiers in Integrative Neuroscience*, *8*, 6. https://doi.org/10.3389/fnint.2014.00006

Bidelman, G. M., & Chung, W. L. (2015). Tone-language speakers show hemispheric specialization and differential cortical processing of contour and interval cues for pitch. *Neuroscience*, *305*, 384–392. https://doi.org/10.1016/j.neuroscience.2015.08.010

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011a). 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. https://doi.org/10.1162/jocn.2009.21362

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011b). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition*, *77*(1), 1–10. https://doi.org/10.1016/j.bandc.2011.07.006

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

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language*, *44*(2), 235–249. *https://doi.org/10.1006/jmla.2000.2750*

Blumenfeld, H. K., & Marian, V. (2013). Parallel language activation and cognitive control during spoken word recognition in bilinguals. *Journal of Cognitive Psychology*, *25*(5), 547–567. *https://doi.org/10.1080/20445911.2013.812093*

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ration of a sampled sound. *Institute of Phonetic Sciences Proceedings, 17,* 97–110.

Boersma, P., & Weenink, D. (2023). Praat: Doing phonetics by computer. [Computer program]. Version 6.3.14, retrieved 4 August 2023 from *http://www.praat.org/*

Bogacka, A. (2004). On the perception of English high vowels by Polish learners of English. *CamLing 2004: Proceedings of the University of Cambridge Second Postgraduate Conference in Language Research*, 43–50.

Bohn, O. S. (1995). Cross language speech production in adults: First language transfer doesn't tell it all. In: W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research,* pp. 279–304. York Press.

Bokander, L. (2020). Language Aptitude and Crosslinguistic Influence in Initial L2 Learning. *Journal of the European Second Language Association*, *4*(1), 35–44. *https://doi.org/10.22599/jesla.69*

Bondarko, L. V. (2005). Phonetic and phonological aspects of the opposition of 'soft' and 'hard' consonants in the modern Russian language. *Speech Communication*, *47*(1–2), 7–14. *https://doi.org/10.1016/j.specom.2005.03.012*

Bradley, E. D. (2012). Tone language experience enhances sensitivity to melodic contour. *LSA Annual Meeting Extended Abstracts*, *3*, 40. *https://doi.org/10.3765/exabs.v0i0.612*

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. *https://doi.org/10.1121/1.427952*

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299–2310. *https://doi.org/10.1121/1.418276*

Braun, B., & Johnson, E. K. (2011). Question or tone 2? How language experience and linguistic function guide pitch processing. *Journal of Phonetics*, *39*(4), 585–594. *https://doi.org/10.1016/j.wocn.2011.06.002*

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7–9), 1044–1098. *https://doi.org/10.1080/01690965.2010.504378*

Brodbeck, C., & Simon, J. Z. (2022). Cortical tracking of voice pitch in the presence of multiple speakers depends on selective attention. *Frontiers in Neuroscience, 16*. *https://doi.org/10.3389/fnins.2022.828546*

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803-809.e3. *https://doi.org/10.1016/j.cub.2018.01.080*

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., & Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, *11*(1), 4963. *https://doi.org/10.1038/s41598-021-84597-9*

Broderick, M. P., & Lalor, E. C. (2020). Co-existence of prediction and error signals in electrophysiological responses to natural speech. [Preprint] bioRxiv. *https://doi.org/10.1101/2020.11.20.391227*

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modelling. *The R Journal*, *9*(2), 378–400. *https://doi.org/10.32614/RJ-2017-066*

Caprini, F., Zhao, S., Chait, M., Agus, T., Pomper, U., Tierney, A., & Dick, F. (2021). Generalization of auditory expertise in audio engineers and instrumental musicians. [Preprint] PsyArXiv. *https://doi.org/10.31234/osf.io/7fg5h*

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81–105. *https://doi.org/10.1016/j.cognition.2014.12.005*

Cason, N., & Schön, D. (2012). Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia*, *50*(11), 2652–2658. *https://doi.org/10.1016/j.neuropsychologia.2012.07.018*

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics*, *34*(3), 372–387. *https://doi.org/10.1016/j.wocn.2005.08.003*

Chait, M., De Cheveigné, A., Poeppel, D., & Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia*, *48*(11), 3262–3271. *https://doi.org/10.1016/j.neuropsychologia.2010.07.007*

Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., & Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. *Nature Communications*, *8*(1), 15027. https://doi.org/10.1038/ncomms15027

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, *5*, 825. https://doi.org/10.3389/fpsyg.2014.00825

Chandrasekaran, B., Krishnan, A., & Gandour, J. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language*, *108*(1), 1–9. https://doi.org/10.1016/j.bandl.2008.02.001

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456–465. https://doi.org/10.1121/1.3445785

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

Cheng, F.-Y., Xu, C., Gold, L., & Smith, S. (2021). Rapid enhancement of subcortical neural responses to sine-wave speech. *Frontiers in Neuroscience*, *15*, 747303. https://doi.org/10.3389/fnins.2021.747303

Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences*, *3*(4), 923–940. https://doi.org/10.3390/brainsci3020923

Choi, W. (2020). The selectivity of musical advantage: Musicians exhibit perceptual advantage for some but not all Cantonese tones. *Music Perception, 37*(5), 423–434. https://doi.org/10.1525/mp.2020.37.5.423

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1468–1479. https://doi.org/10.1044/2014_JSLHR-L-13-0279

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *The Journal of the Acoustical Society of America*, *144*, EL172–EL177. https://doi.org/10.1121/1.5052025

Cobb, T. (2012). Compleat lexical tutor for data-driven language learning on the web. https://www.lextutor.ca/

Cohen, Y. E., & Knudsen, E. I. (1999). Maps versus clusters: Different representations of auditory space in the midbrain and forebrain. *Trends in Neurosciences*, *22*(3), 128–135. https://doi.org/10.1016/S0166-2236(98)01295-8

Corey, R. M., Jones, U., & Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America*, *148*(4), 2371–2375. https://doi.org/10.1121/10.0002279

Corretge, R. (2023). Praat Vocal Toolkit. https://www.praatvocaltoolkit.com/index.html

Costa-Faidella, J., Sussman, E. S., & Escera, C. (2017). Selective entrainment of brain oscillations drives auditory perceptual organization. *NeuroImage*, *159*, 195–206. https://doi.org/10.1016/j.neuroimage.2017.07.056

Creel, S. C., Weng, M., Fu, G., Heyman, G. D., & Lee, K. (2018). Speaking a tone language enhances musical pitch perception in 3–5-year-olds. *Developmental Science*, *21*(1). https://doi.org/10.1111/desc.12503

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*(2). https://doi.org/10.18637/jss.v034.i02

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research. *Frontiers in Neuroscience*, *15*, 705621. https://doi.org/10.3389/fnins.2021.705621

Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching*, *55*(4), 470–489. https://doi.org/10.1017/S0261444821000537

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In: G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 105–121). The MIT Press.

Cutler, A. (2015a). Representation of second language phonology. *Applied Psycholinguistics*, *36*(1), 115–128. https://doi.org/10.1017/S0142716414000459

Cutler, A. (2015b). Lexical stress in English pronunciation. In: M. Reed & J. M. Lewis (Eds.) *The Handbook of English Pronunciation.* Wiley.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Da Costa, S., Van Der Zwaag, W., Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning in to sound: Frequency-selective attentional filter in human primary auditory cortex. *The Journal of Neuroscience*, *33*(5), 1858–1863. https://doi.org/10.1523/JNEUROSCI.4405-12.2013

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. *https://doi.org/10.1016/j.lindif.2015.04.005*

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924-1937.e9. *https://doi.org/10.1016/j.cub.2019.04.067*

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. *https://doi.org/10.1016/j.heares.2007.01.014*

Davis, S., & Van Summers, W. (1989). Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics*, *17*(4), 339–353. *https://doi.org/10.1016/S0095-4470(19)30449-8*

Degerman, A., Rinne, T., Salmi, J., Salonen, O., & Alho, K. (2006). Selective attention to sound location or pitch studied with fMRI. *Brain Research*, *1077*, 123–134.

De Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *The Journal of Neuroscience*, *37*(27), 6539–6557. *https://doi.org/10.1523/JNEUROSCI.3267-16.2017*

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. *https://doi.org/10.1017/S0142716413000210*

de Looze, C., & Rauzy, S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: Methods and potential. *Interspeech 2011 Proceedings*.

Demuth, K., (2007). Acquisition at the prosody-morphology interface. *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition North America*, 84–91.

De Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, *96*(4), 2037–2047. *https://doi.org/10.1121/1.410145*

Derakhshan, A., & Karimi, E. (2015). The interference of first language and second language acquisition. *Theory and Practice in Language Studies*, *5*(10), 2112. *https://doi.org/10.17507/tpls.0510.19*

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favour of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410.

Desai, M., Field, A. M., & Hamilton, L. S. (2023). Dataset size considerations for robust acoustic and phonetic speech encoding models in EEG. *Frontiers in Human Neuroscience*, *16*, 1001171. https://doi.org/10.3389/fnhum.2022.1001171

Dick, F. K., Lehet, M. I., Callaghan, M. F., Keller, T. A., Sereno, M. I., & Holt, L. L. (2017). Extensive tonotopic mapping across auditory cortex is recapitulated by spectrally directed attention and systematically related to cortical myeloarchitecture. *The Journal of Neuroscience*, *37*(50), 12187–12201. https://doi.org/10.1523/JNEUROSCI.1436-17.2017

Di Liberto, G. M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. A., & Mesgarani, N. (2021). Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage*, *227*, 117586. https://doi.org/10.1016/j.neuroimage.2020.117586

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Domahs, U., Knaus, J., Orzechowska, P., & Wiese, R. (2012). Stress “deafness” in a language with fixed word stress: An ERP study on Polish. *Frontiers in Psychology*, *3*, 439. https://doi.org/10.3389/fpsyg.2012.00439

Drennan, D. P., & Lalor, E. C. (2019). Cortical tracking of complex sound envelopes: Modeling the changes in response with intensity. *Eneuro*, *6*(3). https://doi.org/10.1523/ENEURO.0082-19.2019

Duanmu, S. (1994). Against contour tone units. *Linguistic Inquiry*, *25*(4), 555–608.

Durojaye, C., Knowles, K. L., Patten, K. J., Garcia, M. J., & McBeath, M. K. (2021). When music speaks: An acoustic study of the speech surrogacy of the Nigerian Dùndún talking drum. *Frontiers in Communication*, *6*, 652690. https://doi.org/10.3389/fcomm.2021.652690

Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music*, *34*(3), 337–371. https://doi.org/10.1177/0305735606064842

Elhilali, M., Xiang, J., Shamma, S. A., & Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology*, *7*(6), e1000129. https://doi.org/10.1371/journal.pbio.1000129

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience*, *10*(5), 590–604. https://doi.org/10.1162/089892998562997

Escudero, P. (2007). Second-language phonology: the role pf perception. In: M. C. Pennington (Ed.), *Phonology in context* (pp. 109–134). Springer.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465. https://doi.org/10.1016/j.wocn.2009.07.006

Faris, M. M., Best, C. T., & Tyler, M. D. (2016). An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *The Journal of the Acoustical Society of America*, *139*(1), EL1–EL5. https://doi.org/10.1121/1.4939608

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *The Journal of the Acoustical Society of America*, *97*(3), 1893–1904. https://doi.org/10.1121/1.412063

Feng, J., Tao, S., Wu, X., Alsbury, K., & Liu, C. (2019). The effects of amplitude and duration on the perception of English statements vs questions for native English and Chinese listeners. *The Journal of the Acoustical Society of America*, *145*(5), EL449–EL455. https://doi.org/10.1121/1.5109046

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.2307/2336755

Flege, J. E. (1987). The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*(1), 47–65. https://doi.org/10.1016/S0095-4470(19)30537-6

Flege, J. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*, 425–442.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., & Hillenbrand, J. (1986). Differential use of temporal cues to the /s/–/z/ contrast by native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *79*(2), 508–517. https://doi.org/10.1121/1.393538

Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, *23*(4), 527–552. https://doi.org/10.1017/S0272263101004041

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106,* 2973–2987. https://doi.org/10.1121/1.428116

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech*, *38*(1), 25–55. https://doi.org/10.1177/002383099503800102

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory and Language*, *41*(1), 78–104. https://doi.org/10.1006/jmla.1999.2638

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668–1680. https://doi.org/10.3758/BF03212164

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268–294. https://doi.org/10.1016/j.wocn.2007.06.005

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. https://doi.org/10.1121/1.2945161

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366. https://doi.org/10.1037/0096-1523.28.2.349

Francis, A. L., & Nusbaum, H. C. (2009). Effects of intelligibility on working memory demand for speech perception. *Attention, Perception, & Psychophysics*, *71*(6), 1360–1374. https://doi.org/10.3758/APP.71.6.1360

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*(2), 349–356. https://doi.org/10.1017/S027226310426209X

Fuhrmeister, P., & Myers, E. B. (2021). Structural neural correlates of individual differences in categorical perception. *Brain and Language*, *215*, 104919. https://doi.org/10.1016/j.bandl.2021.104919

Gelman, A., Su, Y.S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., Zheng, T., & Dorie, V. (2022). Package 'arm'. https://cran.r-project.org/package=arm

Georgiou, G. P. (2021a). Effects of phonetic training on the discrimination of second language sounds by learners with naturalistic access to the second language. *Journal of Psycholinguistic Research*, *50*(3), 707–721. https://doi.org/10.1007/s10936-021-09774-3

Georgiou, G. P. (2021b). Toward a new model for speech perception: The Universal Perceptual Model (UPM) of second language. *Cognitive Processing*, *22*(2), 277–289. https://doi.org/10.1007/s10339-021-01017-6

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, *14*, 29–56. https://doi.org/10.1146/annurev.ps.14.020163.000333

Giuliano, R. J., Pfordresher, P. Q., Stanley, E. M., Narayana, S., & Wicha, N. Y. Y. (2011). Native experience with a tone language enhances pitch discrimination and the timing of neural responses to pitch Ccange. *Frontiers in Psychology*, *2*, 146. https://doi.org/10.3389/fpsyg.2011.00146

Goad, H., & White, L. (2019). Prosodic effects on L2 grammars. *Linguistic Approaches to Bilingualism*, *9*(6), 769–808. https://doi.org/10.1075/lab.19043.goa

Godfroid, A., Loewen, S., Jung, S., Park, J.-H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgements measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*(2), 269–297. https://doi.org/10.1017/S0272263114000850

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200.

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 20170007. https://doi.org/10.1515/lingvan-2017-0007

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, *25*, 1–42.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., & McAuley, J. D. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental Science*, *18*(4), 635–644. https://doi.org/10.1111/desc.12230

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–343. https://doi.org/10.1177/0267658312461497

Guenther, F. H., Husain, F. T., Cohen, M. A., Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *The Journal of the Acoustical Society of America*, *106*(5), 2900–2912. https://doi.org/10.1121/1.428112

Guion, S. G., & Pederson, E.(2007). Investigating the role of attention in phonetic learning. In: O. S. Bohn, & M. Munro (Eds.) *Second-language speech learning: The role of language experience in speech perception and production: A festschrift in honor of James E. Flege* (pp. 57–77). Amsterdam: Benjamins.

Gussenhoven, C. (2008). Types of focus in English. In: C. Lee, M. Gordon, & D. Buring (Eds.) *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation* (pp. 83–100). Springer.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*(2B), 613–617. https://doi.org/10.1121/1.1911936

Hannon, E. E., Snyder, J. S., Eerola, T., & Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 956–974. https://doi.org/10.1037/0096-1523.30.5.956

Hansen, N. C., Højlund, A., Møller, C., Pearce, M., & Vuust, P. (2022). Musicians show more integrated neural processing of contextually relevant acoustic features. *Frontiers in Neuroscience*, *16*, 907540. https://doi.org/10.3389/fnins.2022.907540

Hao, W. (2023). A comparative study of Chinese and Western music. *Highlights in Art and Design*, *3*(1), 80–82.

Hao, Y.-C. (2018). Second language perception of Mandarin vowels and tones. *Language and Speech*, *61*(1), 1235–152.

Harding, E. E., Sammler, D., Henry, M. J., Large, E. W., & Kotz, S. A. (2019). Cortical tracking of rhythm in music and speech. *NeuroImage*, *185*, 96–101. https://doi.org/10.1016/j.neuroimage.2018.10.037

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88. https://doi.org/10.1016/j.cognition.2019.03.011

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469–479. https://doi.org/10.1121/1.3021295

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*, 46–51. https://doi.org/10.3758/BF03203619

Hazan, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Perception & Psychophysics*, *49*(2), 187–200. https://doi.org/10.3758/BF03205038

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00035

Heinze, G., Ploner, M., Dunkler, D., Southworth, H., Jiricka, L., & Steiner, G. (2022). Package 'logistif'. *https://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/firth-correction/*

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86. *https://doi.org/10.1016/j.jml.2016.01.001*

Hill, K. T., & Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex*, *20*, 583–590.

Hillebrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*, 3099–3111. *https://doi.org/10.1121/1.411872*

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*(4108), 177–180. *https://doi.org/10.1126/science.182.4108.177*

Højen, A. (n.d.). Improvement in Young Adults' Second-language Pronunciation after Short-term Immersion. In: A. M. Nyvad, M. Hejna, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds). *A Sound Approach to Language Matters*, 543–560.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, *42*(3–4). *https://doi.org/10.1515/FLIN.2008.331*

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071. *https://doi.org/10.1121/1.2188377*

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, *72*(5), 1218–1227. *https://doi.org/10.3758/APP.72.5.1218*

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hearing Research*, *366*, 50–64. *https://doi.org/10.1016/j.heares.2018.06.014*

Hove, M. J., Sutherland, M. E., & Krumhansl, C. L. (2010). Ethnicity effects in relative pitch. *Psychonomic Bulletin & Review*, *17*(3), 310–316. *https://doi.org/10.3758/PBR.17.3.310*

Huang, B. H., & Jun, S.-A. (2011). The effect of age on the acquisition of second language prosody. *Language and Speech*, *54*(3), 387–414. *https://doi.org/10.1177/0023830911402599*

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089–1115. https://doi.org/10.1017/S0142716420000193

Hutka, S., Bidelman, G. M., & Moreno, S. (2015). Pitch expertise is not created equal: Cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia*, *71*, 52–63. https://doi.org/10.1016/j.neuropsychologia.2015.03.019

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950–3964. https://doi.org/10.1121/1.4765076

Ihara, A. S., Matsumoto, A., Ojima, S., Katayama, J., Nakamura, K., Yokota, Y., Watanabe, H., & Naruse, Y. (2021). Prediction of second language proficiency based on electroencephalographic signals measured while listening to natural speech. *Frontiers in Human Neuroscience*, *15*, 665809. https://doi.org/10.3389/fnhum.2021.665809

Ingvalson, E. M., Ettlinger, M., & Wong, P. C. M. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism*, *18*(1), 35–47. https://doi.org/10.1177/1367006912456586

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics*, *39*(4), 571–584. https://doi.org/10.1016/j.wocn.2011.03.003

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562.

Iverson, P., & Kuhl, P. K. (1994). Tests of the perceptual magnet effect for American English /r/ and /l/. *The Journal of the Acoustical Society of America*, *95*(5_Supplement), 2976–2976. https://doi.org/10.1121/1.408983

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. https://doi.org/10.1016/S0010-0277(02)00198-1

Jasmin, K., Dick, F., Holt, L. L., & Tierney, A. (2020). Tailored perception: Individuals' speech and music perception strategies fit their perceptual abilities. *Journal of Experimental Psychology: General*, *149*(5), 914–934. https://doi.org/10.1037/xge0000688

Jasmin, K., Dick, F., Stewart, L., & Tierney, A. T. (2020). Altered functional connectivity during speech perception in congenital amusia. *ELife*, *9*, e53539. https://doi.org/10.7554/eLife.53539

Jasmin, K., Sun, H., & Tierney, A. T. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition*, *206*, 104481. https://doi.org/10.1016/j.cognition.2020.104481

Jasmin, K., Tierney, A., Obasih, C., & Holt, L. (2023). Short-term perceptual reweighting in suprasegmental categorization. *Psychonomic Bulletin & Review*, *30*(1), 373–382. https://doi.org/10.3758/s13423-022-02146-5

Jenkinson, M. & Chappell, M. (2018). *Introduction to neuroimaging analysis.* Oxford University Press.

Jeong, H., Li, P., Suzuki, W., Sugiura, M., & Kawashima, R. (2021). Neural mechanisms of language learning from social contexts. *Brain and Language*, *212*, 104874. https://doi.org/10.1016/j.bandl.2020.104874

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., & Yang, Y. (2010). Processing melodic contour and speech intonation in congenital amusics with Mandarin Chinese. *Neuropsychologia*, *48*(9), 2630–2639. https://doi.org/10.1016/j.neuropsychologia.2010.05.009

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15–24. https://doi.org/10.1016/j.bandl.2019.02.004

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591. https://doi.org/10.1017/S0142716417000534

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kamiyama, T. (2011). Pronunciation of French vowels by Japanese speakers learning French as a foreign language: Back and front rounded vowels /u y ø/. *Phonological Studies Phonological Society of Japan, 14*, 97–108.

Kamiyama, T., & Vaissière, J. (2009). Perception and production of French close and close-mid rounded vowels by Japanese-speaking learners. *Acquisition et Interaction En Langue Étrangère*, *Aile... Lia, 2*, 9–41. https://doi.org/10.4000/aile.4533

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594–1611. *https://doi.org/10.1037/xhp0000410*

Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In: Divenyi, P. (Ed.), *Speech Separation by Humans and Machines*. Springer.

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*. *https://doi.org/10.3389/fnhum.2014.00327*

Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050–1060. *https://doi.org/10.1121/1.4730893*

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435. *https://doi.org/10.1121/1.2743154*

Kim, D., Clayards, M., & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. *Linguistics Vanguard*, *3*(1), 20160025. *https://doi.org/10.1515/lingvan-2016-0025*

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. *https://doi.org/10.1016/j.wocn.2020.100984*

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19–EL25. *https://doi.org/10.1121/1.4807431*

Kim, Y. H., & Hazan, V. (2010). Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates. In: K. Dziubalska-Kolaczyk, M. Wrembel, & M. Kul (Eds.) *New Sounds 2010: Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech.*

Kivisto-de Souza, H., Carlet, A., Julkowska, I. A., & Rato, A. (2017). Vowel inventory size matters: Assessing cue-weighting in L2 vowel perception. *Ilha Desterro*, *70*(3). *https://doi.org/10.5007/2175-8026.2017v70n3p33*

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*, 1208–1221. *https://doi.org/10.1121/1.380986*

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*, 3959–3971. https://doi.org/10.1121/1.2999341

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. https://doi.org/10.1016/j.wocn.2010.08.003

Kong, E. J., & Edwards, J. (2015). Individual differences in L2 learners' perceptual cue weighting patterns. *ICPhS 2015 Proceedings.*

Kong, E. J., & Edwards, J. (2011). Individual differences in speech perception: Evidence from visual analogue scaling and eye-tracking. *ICPhS 2011 Proceedings.*

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57. https://doi.org/10.1016/j.wocn.2016.08.006

Kong, E. J., & Kang, S. (2022). Individual differences in categorical judgment of L2 stops: A link to proficiency and acoustic cue-weighting. *Language and Speech*, *66*(2), 354–380. https://doi.org/10.1177/00238309221108647

Kong, E. J., & Lee, H. (2018). Attentional modulation and individual differences in explaining the changing role of fundamental frequency in Korean laryngeal stop perception. *Language and Speech*, *61*(3), 384–408. https://doi.org/10.1177/0023830917729840

Kong, E. J., & Yoon I. H. (2013). L2 proficiency effect on the acoustic cue-weighting pattern by Korean L2 learners of English: Production and perception of English stops. *Journal of the Korean Society of Speech Sciences*, *5*(4), 81–90. http://dx.doi.org/10.13064/KSSS.2013.5.4.081

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krishnan, A., Bidelman, G. M., & Gandour, J. T. (2010). Neural representation of pitch salience in the human brainstem revealed by psychophysical and electrophysiological indices. *Hearing Research*, *1–2, 60–66.* https://doi.org/10.1016/j.heares.2010.04.016

Krishnan, A., Bidelman, G. M., Smalt, C. J., Ananthakrishnan, S., & Gandour, J. T. (2012). Relationship between brainstem, cortical and behavioural measures relevant to pitch salience in humans. *Neuropsychologia, 12*, 2849–2859. https://doi.org/10.1016/j.neuropsychologia.2012.08.013

Krishnan, A., Swaminathan, J., & Gandour, J. T. (2009). Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *Journal of Cognitive Neuroscience*, *21*(6), 1092–1105. https://doi.org/10.1162/jocn.2009.21077

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168. https://doi.org/10.1016/j.cogbrainres.2005.05.004

Krumbholz, K. (2003). Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. *Cerebral Cortex*, *13*(7), 765–772. https://doi.org/10.1093/cercor/13.7.765

Kuang, J., Chan, M. P. Y., & Rhee, N. (2022). *The effects of syntactic and acoustic cues on the perception of prosodic boundaries*. *Speech Prosody 2022 Proceedings,* 699–703. https://doi.org/10.21437/SpeechProsody.2022-142

Kuang, J., & Cui, A. (2018). Relative cue weighting in production and perception of an ongoing sound change in Southern Yi. *Journal of Phonetics*, *71*, 194–214. https://doi.org/10.1016/j.wocn.2018.09.002

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuhl, P. K. (2000). A new view of language acquisition. *PNAS*, *97*(22), 11850–11857. https://doi.org/10.1073/pnas.97.22.11850

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364

Ladányi, E., Novakovic, M., Boorom, O. A., Aaron, A. S., Scartozzi, A. C., Gustavson, D. E., Nitin, R., Bamikole, P. O., Vaughan, C., Fromboluti, E. K., Schuele, C. M., Camarata, S. M., McAuley, J. D., & Gordon, R. L. (2023a). Using motor tempi to understand rhythm and grammatical skills in developmental language disorder and typical language development. *Neurobiology of Language*, *4*(1), 1–28. https://doi.org/10.1162/nol_a_00082

Laffere, A., Dick, F., Holt, L. L., & Tierney, A. (2021). Attentional modulation of neural entrainment to sound streams in children with and without ADHD. *NeuroImage*, *224*, 117396. https://doi.org/10.1016/j.neuroimage.2020.117396

Laffere, A., Dick, F., & Tierney, A. (2020). Effects of auditory selective attention on neural phase: Individual differences and short-term training. *NeuroImage*, *213*, 116717. https://doi.org/10.1016/j.neuroimage.2020.116717

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics.* Cengage learning.

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language*, *66*(1), 285–306. https://doi.org/10.1016/j.jml.2011.09.004

Lee, A. K. C., Larson, E., Maddox, R. K., & Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hearing Research*, *307*, 111–120. https://doi.org/10.1016/j.heares.2013.06.010

Lee, H., & Jongman, A. (2019). Effects of sound change on the weighting of acoustic cues to the three-way laryngeal stop contrast in Korean: Diachronic and dialectal comparisons. *Language and Speech*, *62*(3), 509–530. https://doi.org/10.1177/0023830918786305

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*(2), 417–441. https://doi.org/10.1162/089976699300016719

Lee, A. K. C., Rajaram, S., Xia, J., Bharadwaj, H. M., Larson, E., Hamalainen, M. S., & Shinn-Cunningham, B. G. (2013). Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. *Frontiers in Neuroscience*, *6*, 190. https://doi.org/10.3389/fnins.2012.00190

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science*, *41*, 885–912. https://doi.org/10.1111/cogs.12413

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328. https://doi.org/10.1016/j.cognition.2020.104328

Lehet, M. I., Fenn, K. M., & Nusbaum, H. C. (2020). Shaping perceptual learning of synthetic speech through feedback. *Psychonomic Bulletin & Review*, *27*(5), 1043–1051. https://doi.org/10.3758/s13423-020-01743-6

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from the perception of Southern British English vowels by native speakers of Greek and Japanese. *Phonetica*, *66*(3), 169–187. https://doi.org/10.1159/000235659

Lengeris, A. (2012). Prosody and second language teaching: Lessons from L2 speech perception and production research. In J. Romero-Trillo (Ed.), *Pragmatics and Prosody in English Language Teaching*, 25–40. Springer Netherlands. https://doi.org/10.1007/978-94-007-3883-6_3

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(4), 431–461.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, *243*, 489–494.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*, 467–477. https://doi.org/10.1121/1.1912375

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, *127*(4), 2599–2610. https://doi.org/10.1121/1.3295689

Li, P., & Jeong, H. (2020). The social brain of language: Grounding second language learning in social interaction. *NPJ Science of Learning*, *5*(8), https://doi.org/10.1038/s41539-020-0068-7

Li, P., Legault, J., & Litcofsky, K. A. (2014). Neuroplasticity as a function of second language learning: Anatomical changes in the human brain. *Cortex*, *58*, 301–324. https://doi.org/10.1016/j.cortex.2014.05.001

Li, Y., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. *Nature Communications*, *12*(1), 1161. https://doi.org/10.1038/s41467-021-21430-x

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, *29*(1), 3–11.

Lim, S., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Liu, D., Wang, S., Gao, Q., Dong, R., Fu, X., Pugh, E., & Hu, J. (2020). Learning a second language in adulthood changes subcortical neural encoding. *Neural Plasticity*, *2020*, 1–9. https://doi.org/10.1155/2020/8836161

Liu, F., Maggu, A. R., Lau, J. C. Y., & Wong, P. C. M. (2015). Brainstem encoding of speech and musical stimuli in congenital amusia: Evidence from Cantonese speakers. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.01029

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain*, *133*(6), 1682–1693. https://doi.org/10.1093/brain/awq089

Liu, J., Hilton, C. B., Bergelson, E., & Mehr, S. A. (2023). Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Biology*, *33*(10), 1916-1925.e4. https://doi.org/10.1016/j.cub.2023.03.067

Liu, L., Ong, J. H., Tuninetti, A., & Escudero, P. (2018). One way or another: Evidence for perceptual asymmetry in pre-attentive learning of non-native contrasts. *Frontiers in Psychology*, *9,* 162. https://doi.org/10.3389/fpsyg.2018.00162

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798. https://doi.org/10.1037/xhp0000092

Llanos, F., Dmitrieva, O., Shultz, A., & Francis, A. L. (2013). Auditory enhancement and second language experience in Spanish and English weighting of secondary voicing cues. *The Journal of the Acoustical Society of America*, *134*(3), 2213–2224. https://doi.org/10.1121/1.4817845

Loebach, J. L., & Pisoni, D. P. (2008). Perceptual learning of spectrally degraded speech and environmental sounds. *The Journal of the Acoustical Society of America, 123*, 1126–1139. https://doi.org/10.1121/1.2823453

Lunden, A. (2017). Duration, vowel quality, and the rhythmic pattern of English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 27. https://doi.org/10.5334/labphon.37

MacKay, I. R. A., Flege, J. E., Piske, T., & Schirru, C. (2001). Category restructuring during second-language speech acquisition. *The Journal of the Acoustical Society of America, 110*, 516–528. https://doi.org/10.1121/1.1377287

Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C. S., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., Kenyon, K., Caverlé, M. W., Schultz, B. G., & Vogel, A. P. (2020). Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *The Journal of the Acoustical Society of America*, *148*(6), 3562–3568. https://doi.org/10.1121/10.0002873

Mankel, K., Barber, J., & Bidelman, G. M. (2020). Auditory categorical processing for speech is modulated by inherent musical listening skills. *NeuroReport*, *31*(2), 162–166. https://doi.org/10.1097/WNR.0000000000001369

Marslen-Wilson, W. D., Tyler, L. K., Warren, P., Grenier, P., & Lee, C. S. (1992). Prosodic effects in minimal attachment. *The Quarterly Journal of Experimental Psychology Section A*, *45*(1), 73–87. https://doi.org/10.1080/14640749208401316

Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *The Journal of the Acoustical Society of America*, *60*(3), 704–717. https://doi.org/10.1121/1.381143

Massaro, D. W., & Cohen, M. M. (1977). Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics*, *22*(4), 373–382. https://doi.org/10.3758/BF03199703

Matthews, T. E., Thibodeau, J. N. L., Gunther, B. P., & Penhune, V. B. (2016). The impact of instrument-specific musical training on rhythm perception and production. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00069

May, L., & Werker, J.F. (2014). Can a click be a word? Infants' learning of non-native words. *Infancy*, *19*(3), 281–300. https://doi.org/10.1111/infa.12048

Mayo, C., Scobbie, J. M., Hewlett, N., & Waters, D. (2003). The influence of phonemic awareness development on acoustic cue weighting strategies in children's speech perception. *Journal of Speech, Language, and Hearing Research*, *46*(5), 1184–1196. https://doi.org/10.1044/1092-4388(2003/092)

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, *30*(2), 229–258. https://doi.org/10.1006/jpho.2002.0174

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 89–108. https://doi.org/10.3758/CABN.2.2.89

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McConkey-Robbins, A., Green, J. E., & Waltzman, S. B. (2004). Bilingual Oral Language Proficiency in Children With Cochlear Implants. *Archives of Otolaryngology–Head & Neck Surgery*, *130*(5), 644. https://doi.org/10.1001/archotol.130.5.644

McKercher, D. A. (2018). Overgeneralization. In J. I. Liontas & M. DelliCarpini (Eds.), *The TESOL Encyclopedia of English Language Teaching*, 1–6. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0087

McQueen, J. (1996). Phonetic categorisation. *Language and Cognitive Processes*, *11*(6), 655–664. https://doi.org/10.1080/016909696387060

Mercier, J., Pivneva, I., & Titone, D. (2014). Individual differences in inhibitory control relate to bilingual spoken word processing. *Bilingualism: Language and Cognition*, *17*(1), 89–117. https://doi.org/10.1017/S1366728913000084

Meyer, J., Dentel, L., & Meunier, F. (2017). Categorization of natural whistled vowels by naïve listeners of different language background. *Frontiers in Psychology*, *08*. https://doi.org/10.3389/fpsyg.2017.00025

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621. https://doi.org/10.1111/ejn.13748

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1–2), 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Mickan, A>, McQueen, J. M., Brehm, L., & Lemhofer, K. (2022). Individual differences in foreign language attrition: A 6-month longitudinal investigation after study abroad. *Language, Cognition, and Neuroscience, 38*(1), 11–39. https://doi.org/10.1080/23273798.2022.2074479

Micó, P. (2023). Continuous Dynamic Time Warping (*https://www.mathworks.com/matlabcentral/fileexchange/16350-continuous-dynamic-time-warping*), MATLAB Central File Exchange. Retrieved July 13, 2023.

Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. *https://doi.org/10.1016/j.lingua.2011.04.006*

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. *https://doi.org/10.1016/j.jml.2007.11.006*

Miyake, A., & Friedman, N.P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In: A.F. Healy, & L.E. Bourne (Eds.), *Foreign language learning. Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum Associates.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331–340. *https://doi.org/10.3758/BF03211209*

Mokari, P. G., & Werner, S. (2017) Individual variability in cue weighting for first-language words. *Loquens*, *4*(2), e044. *https://doi.org/10.3989/loquens.2017.044*

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. *https://doi.org/10.1017/S0272263101004016*

Näätänen, R. (1988). Implications of ERP data for psychological theories of attention. *Biological Psychology*, *26*, 117–163.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*(6615), 432–434. *https://doi.org/10.1038/385432a0*

Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, *44*(2), 580–605. *https://doi.org/10.1017/S0272263121000371*

Nagle, C. L., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility, and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, *6*(3), 329–351. *https://doi.org/10.1075/jslp.20009.nag*

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of Neuroscience*, *31*(5), 1704–1714. *https://doi.org/10.1523/JNEUROSCI.4853-10.2011*

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nguyen, D. D., McCabe, P., Thomas, D., Purcell, A., Doble, M., Novakovic, D., Chacon, A., & Madill, C. (2021). Acoustic voice characteristics with and without wearing a facemask. *Scientific Reports*, *11*(1), 5651. https://doi.org/10.1038/s41598-021-85130-8

Nguyễn, T. A.-T., Ingram, C. L. J., & 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

Norman-Haignere, S. V., Albouy, P., Caclin, A., McDermott, J. H., Kanwisher, N. G., & Tillmann, B. (2016). Pitch-responsive cortical regions in congenital amusia. *The Journal of Neuroscience*, *36*(10), 2986–2994. https://doi.org/10.1523/JNEUROSCI.2705-15.2016

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, *19*, 131–150. https://doi.org/10.3758/bf03197110

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In: A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes* (pp. 149–167). Lawrence Erlbaum Associates, Inc.

Nosofsky, R. M., & Hu, M. (2022). Category structure and region-specific selective attention. *Memory & Cognition*, *51*(4), 915–929. https://doi.org/10.3758/s13421-022-01365-4

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*, *31*(28), 10234–10240. https://doi.org/10.1523/JNEUROSCI.0411-11.2011

Nusbaum, H. (1997). Talker normalization: Phonetic constancy as a cognitive process. In: K. A. Johnson & J. W Mullenix (Eds.) *Talker variability and speech processing,* 109–132. Academic Press.

Nusbaum, H. C., & Goodman, J. C. (1994). Learning to hear speech as spoken language. In: J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 299–338). The MIT Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In: Y. Tohkura, E. Vaitkiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure.* IOS Press

Obasih, C. O., Luthra, S., Dick, F., & Holt, L. L. (2023). Auditory category learning is robust across training regimes. *Cognition*, *237*, 105467. https://doi.org/10.1016/j.cognition.2023.105467

Oleson, J. J., Cavanaugh, J. E., Tomblin, J. B., Walker, E., & Dunn, C. (2016). Combining growth curves when a longitudinal study switches measurement tools. *Statistical Methods in Medical Research*, *25*(6), 2925–2938. https://doi.org/10.1177/0962280214534588

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Ordin, M., Polyanskaya, L., Laka, I., & Nespor, M. (2017). Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory & Cognition*, *45*(5), 863–876. https://doi.org/10.3758/s13421-017-0700-9

Ou, J., & Law, S.-P. (2017). Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Attention, Perception, & Psychophysics*, *79*(3), 945–963. https://doi.org/10.3758/s13414-017-1283-z

Ou, J., Law, S.-P., & Fung, R. (2015). Relationship between individual differences in speech processing and cognitive functions. *Psychonomic Bulletin & Review*, *22*(6), 1725–1732. https://doi.org/10.3758/s13423-015-0839-y

Ou, J., Yu, A. C. L., & Xiang, M. (2021). Individual differences in categorization gradience as predicted by online processing of phonetic cues during spoken word recognition: evidence from eye movements. *Cognitive Science*, *45*(3). https://doi.org/10.1111/cogs.12948

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from first language Processing. *Language Learning*, *66*(4), 900–944. https://doi.org/10.1111/lang.12168

Palmer, C., & Krumhansl, C. L. (1987). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, *41*(6), 505–518. https://doi.org/10.3758/BF03210485

Paltoglou, A. E., Sumner, C. J., & Hall, D. A. (2009). Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention. *Hearing Research*, *257*(1–2), 106–118. https://doi.org/10.1016/j.heares.2009.08.007

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

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00057

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic speech adaptations to pitch perturbation in running speech. *Journal of Speech, Langauge, and Hearing Research*, *54*(4), 1051–1059. https://doi.org/10.1044/1092-4388(2010/10-0162)

Peng, S.-C., Lu, N., & Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiology and Neurotology*, *14*(5), 327–337. https://doi.org/10.1159/000212112

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In: C. Gussenhoven & N. Warner (Eds.) *Laboratory Phonology 7*. De Gruyter Mouton.

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, *38*(3), 422–430. https://doi.org/10.1016/j.wocn.2010.04.001

Peretz, I. (2016). Neurobiology of congenital amusia. *Trends in Cognitive Sciences*, *20*(11), P857–867. https://doi.org/10.1016/j.tics.2016.09.002

Peretz, I., Nguyen, S., & Cummings, S. (2011). Tone language fluency impairs pitch discrimination. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00145

Perez, M. M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, *42*, 749–773. https://doi.org/10.1017/S0272263119000706

Petroni, F., & Serva, M. (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and Its Applications*, *389*(11), 2280–2283. https://doi.org/10.1016/j.physa.2010.02.004

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics*, *71*(6), 1385–1398. https://doi.org/10.3758/APP.71.6.1385

Piske, T., Flege, J. E., MacKay, I. R. A., & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica*, *59*(1), 49–71. https://doi.org/10.1159/000056205

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In: J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 121–166). The MIT Press.

Plack, C. J., Barker, D., & Hall, D. A. (2014). Pitch coding and pitch processing in the human brain. *Hearing Research*, *307*, 53–64. https://doi.org/10.1016/j.heares.2013.07.020

Plag, I., Kunter, G., & Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, *39*(3), 362–374. https://doi.org/10.1016/j.wocn.2011.03.004

Platt, C. J., & Haykin, S. (n.d.). *An Information-Maximization Approach to Blind Separation and Blind Deconvolution*.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, *89*(6), 2961–2977. https://doi.org/10.1121/1.400734

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences*, *106*(26), 10598–10602. https://doi.org/10.1073/pnas.0904134106

Press, C., Yon, D., & Heyes, C. (2022). Building better theories. *Current Biology*, *32*(1), R13–R17. https://doi.org/10.1016/j.cub.2021.11.027

Price, C. N., & Moncrieff, D. (2021). Defining the role of attention in hierarchical auditory processing. *Audiology Research*, *11*(1), 112–128. https://doi.org/10.3390/audiolres11010012

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, *119*(3), 1684–1696. https://doi.org/10.1121/1.2161427

Qin, Z., Zhang, C., & Wang, W. S. (2021). The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *The Journal of the Acoustical Society of America*, *149*(1), 435–446. https://doi.org/10.1121/10.0003330

Reetzke, R., Gnanateja, G. N., & Chandrasekaran, B. (2021). Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain and Language*, *213*, 104891. https://doi.org/10.1016/j.bandl.2020.104891

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech*, *54*(2), 147–165. https://doi.org/10.1177/0023830910397489

Repp, B. H. (1983). Coarticulation in sequences of two nonhomorganic stop consonants: Perceptual and acoustic evidence. *The Journal of the Acoustical Society of America*, *74*(2), 420–427. https://doi.org/10.1121/1.389835

Rinne, T., Särkkä, A., Degerman, A., Schröger, E., & Alho, K. (2006). Two separate mechanisms underlie auditory change detection and involuntary control of attention. *Brain Research*, *1077*(1), 135–143. https://doi.org/10.1016/j.brainres.2006.01.043

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2023). Package 'MASS'. *http://www.stats.ox.ac.uk/pub/MASS4/*

Roark, C. L., & Chandrasekaran, B. (2023). Stable, flexible, common, and distinct behaviors support rule-based and information-integration category learning. *Npj Science of Learning*, *8*(1), 14. *https://doi.org/10.1038/s41539-023-00163-0*

Roark, C. L., & Holt, L. L. (2019). Perceptual dimensions influence auditory category learning. *Attention, Perception, & Psychophysics*, *81*(4), 912–926. *https://doi.org/10.3758/s13414-019-01688-6*

Rodero, E. (2011). Intonation and emotion: Influence of pitch levels and contour type on creating emotions. *Journal of Voice*, *25*(1), e25–e34. *https://doi.org/10.1016/j.jvoice.2010.02.002*

Rogers, C. L., & Dalby, J. (2005). Forced-choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research*, *48*(2), 306–322. *https://doi.org/10.1044/1092-4388(2005/021)*

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In: T. E. Moore (Ed.), *Cognitive Development and Acquisition of Language*, 111–144. Academic Press.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192–233. *https://doi.org/10.1037/0096-3445.104.3.192*

Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D., & Pillow, J. W. (2018). Efficient inference for time-varying behavior during learning. *Advances in Neural Information Processing Systems Proceedings.*

Roy, N. A., Bak, J. H., The International Brain Laboratory, Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in decision-making experiments. *Neuron, 109*, 597–610. *https://doi.org/10.1016/j.neuron.2020.12.004*

Ruan, Y., & Saito, K. (2023). Less precise auditory processing limits instructed L2 speech learning: Communicative focus on phonetic form revisited. *System*, *114*, 103020. *https://doi.org/10.1016/j.system.2023.103020*

Saito, K. (2013). Age effects on late bilingualism: The production development of /ɹ/ by high-proficiency Japanese learners of English. *Journal of Memory and Language*, *69*(4), 546–562. *https://doi.org/10.1016/j.jml.2013.07.003*

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, *65*(3), 563–595. *https://doi.org/10.1111/lang.12120*

Saito, K. (2015). The role of age of acquisition in late second language oral proficiency attainment. *Studies in Second Language Acquisition*, *37*(4), 713–743. *https://doi.org/10.1017/S0272263115000248*

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*, *37*(4), 813–840. *https://doi.org/10.1017/S0142716415000259*

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, *115*, 104168. *https://doi.org/10.1016/j.jml.2020.104168*

Saito, K., Kachlicka, M., Suzukida, Y., Mora-Plaza, I., Ruan, Y., & Tierney, A. (*under review*). Auditory processing as perceptual, cognitive and motoric abilities underlying successful second language acquisition. *Journal of Experimental Psychology: Human Perception & Performance.*

Saito, K., Kachlicka, M., Suzukida, Y., Petrova, K., Lee, B. J., & Tierney, A. (2022). Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning. *Cognition*, *229*, 105236. *https://doi.org/10.1016/j.cognition.2022.105236*

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. *https://doi.org/10.1111/lang.12345*

Saito, K., Sun, H., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. (2022). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, *44*(1), 57–86. *https://doi.org/10.1017/S0272263120000467*

Saito, K., Sun, H., & Tierney, A. (2019). Explicit and implicit aptitude effects on second language speech learning: Scrutinizing segmental and suprasegmental sensitivity and performance via behavioural and neurophysiological measures. *Bilingualism: Language and Cognition*, *22*(5), 1123–1140. *https://doi.org/10.1017/S1366728918000895*

Saito, K., & Tierney, A. (2022). Domain-general auditory processing as a conceptual and measurement framework for second language speech learning aptitude: A test-retest reliability study. *Studies in Second Language Acquisition*, 1–25. *https://doi.org/10.1017/S027226312200047X*

Schepens, J. J., Van Der Slik, F., & Van Hout, R. (2016). L1 and L2 distance effects in learning L3 Dutch. *Language Learning*, *66*(1), 224–256. *https://doi.org/10.1111/lang.12150*

Schepens, J. J., Van Hout, R. W. N. M., & Van Der Slik, F. W. P. (2023). Linguistic dissimilarity increases age-related decline in adult language learning. *Studies in Second Language Acquisition*, *45*(1), 167–188. *https://doi.org/10.1017/S0272263122000067*

Schepens, J., Van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056. https://doi.org/10.1016/j.cognition.2019.104056

Schertz, J., Carbonell, K., & Lotto, A. J. (2019). Language specificity in phonetic cue weighting: Monolingual and bilingual perception of the stop voicing contrast in English and Spanish. *Phonetica*, *77*(3), 186–208. https://doi.org/10.1159/000497278

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*(1), 355–367. https://doi.org/10.3758/s13414-015-0987-1

Schertz, J., & Clare, E. J. (2019). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, *11*(2). https://doi.org/10.1002/wcs.1521

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Schönwiesner, M., & Zatorre, R.J. (2008). Depth electrode recordings show double dissociation between pitch processing in lateral Heschl's gyrus and sound onset processing in medial Heschl's gyrus. *Experimental Brain Research*, 187, 97–105. https://doi.org/10.1007/s00221-008-1286-z

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, *12*(1), 40–72. https://doi.org/10.1177/026765839601200103

Schwartz, G., Aperlinski, G., Jekiel, M., & Malarski, K. (2016). Spectral dynamics in L1 and L2 vowel perception. *Research in Language*, *14*(1). https://doi.org/10.1515/rela-2016-0004

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *The Journal of the Acoustical Society of America*, *71*(4), 996–1007. https://doi.org/10.1121/1.387581

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*, *33*(4), 644–653. https://doi.org/10.3758/BF03195331

Shamloo, F., & Hélie, S. (2020). A study of individual differences in categorization with redundancy. *Journal of Mathematical Psychology*, *99*, 102467. https://doi.org/10.1016/j.jmp.2020.102467

Shao, Y., Saito, K., & Tierney, A. (2023). How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training. *TESOL Quarterly*, *57*(1), 33–63. *https://doi.org/10.1002/tesq.3120*

Shiffrin, M. R., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In: R. L. Goldstone, D. L. Media, & P. G. Schyns (Eds.) *Perceptual learning*. Academic Press.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. *https://doi.org/10.1016/j.tics.2008.02.003*

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Shuai, L., & Elhilali, M. (2014). Task-dependent neural representations of salient events in dynamic auditory scenes. *Frontiers in Neuroscience*, *8*. *https://doi.org/10.3389/fnins.2014.00203*

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1–32. *https://doi.org/10.1016/j.cogpsych.2006.04.002*

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95–EL101. *https://doi.org/10.1121/1.4736711*

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(02), 287–308. *https://doi.org/10.1017/S0952675706000911*

Slater, J., Kraus, N., Woodruff Carr, K., Tierney, A., Azem, A., & Ashley, R. (2018). Speech-in-noise perception is linked to rhythm production skills in adult percussionists and non-musicians. *Language, Cognition and Neuroscience*, *33*(6), 710–717. *https://doi.org/10.1080/23273798.2017.1411960*

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293. *https://doi.org/10.1177/0023830909357156*

Song, J. H., Skoe, E., Wong, P. C. M., & Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, *20*(10), 1892–1902. *https://doi.org/10.1162/jocn.2008.20131*

Stagray, J. R., & Downs, D. (1993). Differential sensitivity for frequency among speakers of a tone and a nontone langauge. *Journal of Chinese Linguistics*, *21*(1), 143–163. *https://www.jstor.org/stable/23756129*

Stamer, M. K., & Vitevitch, M. S. (2012). Phonological similarity influences word learning in adults learning Spanish as a foreign language. *Bilingualism: Language and Cognition*, *15*(3), 490–502. https://doi.org/10.1017/S1366728911000216

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing Research*, *305*, 57–73. https://doi.org/10.1016/j.heares.2013.05.013

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America*, *64*(6), 1582–1592. https://doi.org/10.1121/1.382142

Streeter, L. A., Krauss, R. M., Geller, V., Olson, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, *35*(5), 345–350.

Su, I.-R. (2001). Transfer of sentence processing strategies: A comparison of L2 learners of Chinese andEnglish. *Applied Psycholinguistics*, *22*(1), 83–112. https://doi.org/10.1017/S0142716401001059

Sun, H., Saito, K., & Tierney, A. (2021). A longitudinal investigation of explicit and implicit auditory processing in L2 segmental and suprasegmental acquisition. *Studies in Second Language Acquisition*, *43*(3), 551–573. https://doi.org/10.1017/S0272263120000649

Sussman, E. S. (2017). Auditory scene analysis: An attention perspective. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2989–3000. https://doi.org/10.1044/2017_JSLHR-H-17-0041

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning.* Academic Press.

Swaminathan, J., Krishnan, A., & Gandour, J. T. (2008). Pitch encoding in speech and nonspeech contexts in the human auditory brainstem. *NeuroReport*, *19*(11), 1163–1167. https://doi.org/10.1097/WNR.0b013e3283088d31

Symons, A. E., Dick, F., & Tierney, A. T. (2021). Dimension-selective attention and dimensional salience modulate cortical tracking of acoustic dimensions. *NeuroImage*, *244*, 118544. https://doi.org/10.1016/j.neuroimage.2021.118544

Symons, A. E., Holt, L. L., & Tierney, A. T. (2023). Individual differences in the effects of informational masking on segmental and suprasegmental speech categorization. [Preprint] *PsyArXiv*. https://doi.org/10.31234/osf.io/9eq6b

Symons, A. E., & Tierney, A. T. (2023). Musical experience is linked to enhanced dimension-selective attention to pitch and increased primary weighting during suprasegmental categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001217

Tang, W., Xiong, W., Zhang, Y., Dong, Q., & Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: Behavioral and neural evidence. *Neuropsychologia*, *91*, 247–253. https://doi.org/10.1016/j.neuropsychologia.2016.08.003

Teoh, E. S., Cappelloni, M. S., & Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, *50*(11), 3831–3842. https://doi.org/10.1111/ejn.14510

Tervaniemi, M. (2009). Musicians-same or different? *Annals of the New York Academy of Sciences*, *1169*(1), 151–156. https://doi.org/10.1111/j.1749-6632.2009.04591.x

Tierney, A., Rosen, S., & Dick, F. (2020). Speech-in-speech perception, nonverbal selective attention, and musical training. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 968–979. https://doi.org/10.1037/xlm0000767

Tillmann, B., & Poulin-Charronnat, B. (2010). Auditory expectations for newly acquired structures. *Quarterly Journal of Experimental Psychology*, *63*(8), 1646–1664. https://doi.org/10.1080/17470210903511228

Tillmann, B., Schulze, K., & Foxton, J. M. (2009). Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds. *Brain and Cognition*, *71*(3), 259–264. https://doi.org/10.1016/j.bandc.2009.08.003

Toffanin, P., De Jong, R., Johnson, A., & Martens, S. (2009). Using frequency tagging to quantify attentional deployment in a visual divided attention task. *International Journal of Psychophysiology*, *72*(3), 289–298. https://doi.org/10.1016/j.ijpsycho.2009.01.006

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434-464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Trainor, L. J., McDonald, K. L., & Alain, C. (2002). Automatic and controlled processing of melodic contour and interval information measured by electrical brain activity. *Journal of Cognitive Neuroscience*, *14*(3), 430–442. https://doi.org/10.1162/089892902317361949

Tremblay, A., Broersma, M., & Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingualism: Language and Cognition*, *21*(3), 640–652. https://doi.org/10.1017/S136672891700030X

Tremblay, A., Broersma, M., Coughlin, C. E., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00985

Tremblay, A., Kim, S., Shin, S., & Cho, T. (2021). Re-examining the effect of phonological similarity between the native- and second-language intonational systems in second-language speech segmentation. *Bilingualism: Language and Cognition*, *24*(2), 401–413. https://doi.org/10.1017/S136672892000053X

Trofimovich, P., & Baker, W. (2001). Investigating the nature of cross-language perceptual comparisons: Evidence from production. *The Journal of the Acoustical Society of America*, *109*, 2472. https://doi.org/10.1121/1.4744776

Trofimovich, P., & Baker, W. (2006). Learning second language segmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *Studies in Second Language Acquisition*, *28*(01). https://doi.org/10.1017/S0272263106060013

Turker, S., & Reiterer, S. M. (2021). Brain, musicality, and language aptitude: A complex interplay. *Annual Review of Applied Linguistics*, *41*, 95–107. https://doi.org/10.1017/S0267190520000148

Turker, S., Seither-Preisler, A., & Reiterer, S. M. (2021). Examining individual differences in language learning: A neurocognitive model of language aptitude. *Neurobiology of Language*, 1–27. https://doi.org/10.1162/nol_a_00042

Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In: A. M. Nyvad, M. Hejna, A. Hojen, A. B. Jespersen, & M. H. Sorensen (Eds.) *A sound approach to language matters–In honor of Ocke-Schwen Bohn.* AU Library Scholarly Publishing Services.

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica*, *71*(1), 4–21. https://doi.org/10.1159/000356237

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Urai, A. E., De Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *ELife*, *8*, e46331. https://doi.org/10.7554/eLife.46331

Van De Weijer, J., & Sloos, M. (2014). The four tones of Mandarin Chinese: Representation and acquisition. *Linguistics in the Netherlands*, *31*, 180–191. https://doi.org/10.1075/avt.31.13wei

Van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01000

Van Staden, A., & Purcell, N. (2016). Multi-sensory learning strategies to support spelling development: A case study of second-language learners with auditory processing difficulties. *International Journal on Language, Literature and Culture in Education*, *3*(1), 40–61. https://doi.org/10.1515/llce-2016-0003

Viswanathan, V., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2019). Electroencephalographic signatures of the neural representation of speech during selective attention. *Eneuro*, *6*(5). https://doi.org/10.1523/ENEURO.0057-19.2019

Wang, J., Friedman, D., Ritter, W., & Bersick, M. (2005). ERP correlates of involuntary attention capture by prosodic salience in speech. *Psychophysiology*, *42*(1), 43–55. https://doi.org/10.1111/j.1469-8986.2005.00260.x

Wang, Q. (2008). L2 Stress Perception: The reliance on different acoustic cues. *Speech Prosody 2008 Proceedings.*

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, *65*, 139–150. https://doi.org/10.1016/j.system.2016.12.013

Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics*, *36*(2), 250–267. https://doi.org/10.1016/j.wocn.2007.06.004

Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., & Höhle, B. (2012). How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology*, *3.* https://doi.org/10.3389/fpsyg.2012.00580

Wells, J. (1962). A study of the formants of the pure vowels of British English. *Unpublished MA Thesis.* https://www.phon.ucl.ac.uk/home/wells/formants/index.htm

Werker, J. F., & Polka, L. (1993). Developmental changes in speech perception: New challenges and new directions. *Journal of Phonetics*, *21*(1–2), 83–101. https://doi.org/10.1016/S0095-4470(19)31322-1

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Wiener, S. (2017). Changes in early L2 cue-weighting of non-native speech: Evidence from learners of Mandarin Chinese. *Interspeech 2017 Proceedings*, 1765–1769. https://doi.org/10.21437/Interspeech.2017-289

Wierzchowska, B. (1971). *Wymowa Polska*. Warszawa: Panstwowe Wydawnictwo Naukowe.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wilcox, R. R. & Schonbrodt, F. (2017). WRS: A package of R. R. Wilcox' robust statistics functions. *https://dornsife.usc.edu/labs/rwilcox/software/*

Winn, M. B. (2014). *http://www.mattwinn.com/praat/Make_Duration_Continuum.txt*

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479. *https://doi.org/10.1121/1.3672705*

Woldorff, M. G., & Hillyard, S. A. (1991). Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology*, *79*(3), 170–191. *https://doi.org/10.1016/0013-4694(91)90136-R*

Wong, P. C. M., Ciocca, V., Chan, A. H. D., Ha, L. Y. Y., Tan, L.-H., & Peretz, I. (2012). Effects of culture on musical pitch perception. *PLOS ONE*, *7*(4), e33424. *https://doi.org/10.1371/journal.pone.0033424*

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*, 420–422. *https://www.nature.com/articles/nn1872*

Wu, P., Shi, J., Zhong, Y., Watanabe, S., & Black, A. W. (2021). Cross-lingual transfer for speech processing using acoustic language similarity. [Preprint] arXiv. *http://arxiv.org/abs/2111.01326*

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(9), 913–925. *https://doi.org/10.1037/xhp0001037*

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, *52*(4), 376–392. *https://doi.org/10.3758/BF03206698*

Yang, M., & Sundara, M. (2019). Cue-shifting between acoustic cues: Evidence for directional asymmetry. *Journal of Phonetics*, *75*, 27–42. *https://doi.org/10.1016/j.wocn.2019.04.002*

Yang, X., Shen, X., Li, W., & Yang, Y. (2014). How listeners weight acoustic cues to intonational phrase boundaries. *PLoS ONE*, *9*(7), e102166. *https://doi.org/10.1371/journal.pone.0102166*

Yazawa, K., Whang, J., Kondo, M., & Escudero, P. (2020). Language-dependent cue weighting: An investigation of perception modes in L2 learning. *Second Language Research*, *36*(4), 557–581. *https://doi.org/10.1177/0267658319832645*

Yeni-Komshian, G. H., Flege, J. E., & Liu, H. (1997). Pronunciation proficiency in L1 and L2 among Korean–English bilinguals: The effect of age of arrival in the U.S. *The Journal of the Acoustical Society of America*, *102*(3188). https://doi.org/10.1121/1.420659

Yeni-Komshian, G. H., Flege, J. E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean–English bilinguals. *Bilingualism: Language and Cognition*, *3*(2), 131–149. https://doi.org/10.1017/S1366728900000225

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Yu, V. Y., & Andruski, J. E. (2010). A cross-language study of perception of lexical stress in English. *Journal of Psycholinguistic Research*, *39*(4), 323–344. https://doi.org/10.1007/s10936-009-9142-2

Zaltz, Y., Globerson, E., & Amir, N. (2017). Auditory perceptual abilities are associated with specific auditory experience. *Frontiers in Psychology*, *8*, 2080. https://doi.org/10.3389/fpsyg.2017.02080

Zeng, Z., Mattock, K., Liu, L., Peter, V., Tuninetti, A., & Tsao, F.-M. (2020). Mandarin and English adults' cue-weighting of lexical stress. *Interspeech 2020 Proceedings*, 1624–1628. https://doi.org/10.21437/Interspeech.2020-2612

Zhang, A., Feng, H., Zheng, X., Xu, Z., & Dang, J. (2015). The perception of English vowel contrasts by Chinese EFL learners and native English speakers. *International Congress of Phonetic Sciences 2015 Proceedings.*

Zhang, J., Li, W., Xie, Y., & Cao, W. (2012). A quantitative study on information contribution of prosody phrase boundaries in Chinese speech. *Speech Prosody 2012 Proceedings.*

Zhang, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2020). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music*, *48*(3), 327-462. https://doi.org/10.1177/0305735618804038

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The learning signal in perceptual tuning of speech: Bottom up versus top-down information. *Cognitive Science*, *45*(3). https://doi.org/10.1111/cogs.12947

Zhang, Y., & Francis, A. (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics*, *38*(2), 260–271. https://doi.org/10.1016/j.wocn.2009.11.002

Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, *123*(6), 4498–4513. https://doi.org/10.1121/1.2902165

Zheng, Y., & Samuel, A. G. (2018). The effects of ethnicity, musicianship, and tone language experience on pitch perception. *Quarterly Journal of Experimental Psychology*, *71*(12), 2627–2642. *https://doi.org/10.1177/1747021818757435*

Zou, J., Feng, J., Xu, T., Jin, P., Luo, C., Zhang, J., Pan, X., Chen, F., Zheng, J., & Ding, N. (2019). Auditory and language contributions to neural encoding of speech features in noisy environments. *NeuroImage*, *192,* 666–75. *https://doi.org/10.1016/j.neuroimage.2019.02.047*