Pitch perception in lexical tone and melody

Evan Bradley*, Lawrence University; Wisconsin, United States

Music is frequently compared to language, both structurally and psychologically, but the cognitive relationship between the two is still not well understood. This review examines pitch, a salient acoustic property shared by language and music, in order to evaluate the state of knowledge regarding the effects of musical and linguistic experience on the other domain. Specifically, the linguistic property of lexical tone is discussed in relation to musical melody. Basic facts about lexical tone systems are described, along with factors relevant to their perception. Structural components of melody and their perception are discussed, and parallels are drawn between aspects of melodic and linguistic perception. Cases of interaction between linguistic and musical pitch perception are reviewed, and theoretical perspectives on their relationship are compared. A set of perceptual models are identified that generate hypotheses about shared perceptual properties, which have the potential to further explain and specify the mutual influence between music and language cognition.

Citation: Bradley, E. (2013). Pitch perception in lexical tone and melody. Reviews of Research in Human Learning and Music, 1: epub2013002. doi:10.6022/journal.rrhlm.2013002

Copyright: ©2013 Bradley, E. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Key words: Lexical tone; melody; pitch; learning; music and language

*Email: yevb@udel.edu

It has frequently been argued that language and music share structural and cognitive similarities, and experience in each domain has been shown to affect the other in various ways; however, the scope and consequences of this overlap has yet to be fully defined. Better defining this relationship is of interest to music and language researchers, because it has the potential to further illustrate the composition of the perceptual and cognitive systems underlying music and language. Defining this relationship is also of interest to performers and practitioners, because a better understanding of the perceptual and learning mechanisms involved in musical and linguistic performance can further explain and inform educational practices.

In this review, I summarize the state of research on the music-language relationship, attempting to synthesize principles of the relationship and relate them to more general theories of perceptual learning. Rather than survey the vast range of musical and linguistic structures that have been mentioned in the literature, I instead focus on a salient area of overlap between the two: the perceptual property pitch as found in lexical tone and melody.

First, I describe basic facts about lexical tone systems and factors relevant to their perception, with the goal of providing a resource for music researchers interested in lexical tone. Then, I discuss the perception of pitch in music, with an emphasis on identifying parallels between musical and linguistic structures and perceptual phenomena. Next, I review cases of interaction between linguistic and musical pitch perception and compare theoretical perspectives on their relationship. Finally, I propose a mapping between structural components of tone and melody and provide a hypothesis for avenues of research that have the potential to contribute explanatory power to some observed crossover effects and to predict new effects yet to be observed.

Preliminaries

A few notes on terminology are necessary before beginning, as each field of study (i.e., linguistics, music theory) has its own jargon; the same term may have different meanings in different traditions, or the same phenomena may be called by several names. I will attempt to maintain the traditional
meanings of terms as used in their respective fields, while clarifying overlaps or deviations from standard meanings.

Confusion can result from the use of the terms pitch and tone in different contexts, and distinctions must be made between physical, perceptual, and systematic descriptions of auditory phenomena. Yip (2002) succinctly describes the levels that must be considered in regard to lexical tone: Fundamental frequency (F3) is a physical property of the acoustic signal. Pitch is the perceptual object correlated with frequency (and other acoustic properties) by a listener. McDermott and Oxenham (2008, p. 452) describe pitch as “the perceptual correlate of periodicity in sound,” and this perceptual object is available to multiple cognitive systems, including music (e.g., melody, harmony) and language (e.g., tone, intonation). Lexical tone is an abstract linguistic object—part of a speaker’s phonological and lexical knowledge systems. Musical tone is the analogous musical category—the listener’s knowledge of the mapping of pitch to a musically meaningful unit (e.g., A-sharp, sol). Other clarifications of terminology will be made as they are introduced.

Linguistic tone
Language includes multiple levels of structure, including phonetics and phonology (sound structure and organization), morphology (word structure), syntax (sentence structure), and semantics and pragmatics (meaning). Pitch can play a role in many of these, but this review will focus on lexical tone, which is the use of pitch information to distinguish individual words. Tone is different than intonation, which is the use of pitch to distinguish sentence types (e.g., a question from a declarative statement). An intermediate case is the limited use of pitch to distinguish words in a manner similar to stress, or pitch accent (Maddieson, 2005). Thus, one way of characterizing the degree of lexical pitch-use in a language is as a continuum, from intonation-only languages at the less-tonal end, to full-fledged tone languages (and from those with few to those with many tones) at the more-tonal end.

Phonetic and typological facts
By some estimations, as many as 60–70% of the world’s languages can be classified as tonal (Yip, 2002), but only a portion of these have large tonal inventories (Maddieson, 2005). Tone languages are clustered geographically, with the majority occurring in East and Southeast Asia, Africa, and among the indigenous languages of the Americas (Maddieson, 2005). Most tone languages (up to 80%) contain only relatively level tones (Maddieson, 1978); these are known as register tone languages, and are especially common in Africa. Contour tone languages, common in East and Southeast Asia, include rising, falling, concave (falling–rising), or convex (rising–falling) tones. An example of a register tone language with three tones (Low, Mid, and High) is Yoruba (Niger-Congo, Nigeria); an example of a contour tone language is Mandarin (Sino-Tibetan, China), which has four tones, including a high-level, falling, rising, and concave (down then up, or dipping) tone.

Many register tone languages have a simple binary contrast between low and high level tones, but register tone languages may contrast up to five levels, with three being common. Higher numbers of tones are increasingly rare (Gussenhoven, 2004a; Yip, 2002). Phonetically, tones in register tone languages need not be perfectly flat (i.e., stable), but must be at least flat enough that a level pitch is an acceptable version of the tone (Maddieson, 1978). Rising or falling tones may occur in register tone languages as a result of phonological processes, but they do not have the same phonemic status as rising or falling tones in contour tone languages (Yip, 2002).

Register tones are defined by their relative position within the pitch space used by the speakers. Rather than maximal dispersion of tones within the pitch range of a speaker, such that a language with two tones will have a greater distance between its tones than a language with three tones (with each language having a similar distance between its lowest and highest tones), in most register-tone languages, tones are separated by 2–3 semitones, with the total pitch range increasing for languages with more tones (Maddieson, 1978, 1991).

Contour tones tend to appear only in languages with a relatively large number of tones (Maddieson, 2005), and by some accounts, contour tone languages may have up to 8 or even 13 tones (Patel, 2003; Yip, 2002). Segmental, syllabic, and voice quality correlates make determining the exact number of phonemic categories in a given language difficult (Gussenhoven, 2004a). As is observed among register tone languages, contour tone languages with a greater number of tone contrasts are more rare than are those with fewer. Tones with complex contours (concave or convex) are more rare than are simple rising and falling tones (Yip, 2002). Despite their name, contour tone languages also contain level tones; in fact, if a language contains a phonemic
contour tone, this implies that it has at least one level tone (Maddieson (1978) notes a few possible exceptions; Patel, 2008a). Contour tones are rare in languages with only three tones, suggesting that contour tones may result when an upper bound on level tones is reached (Patel, 2008a; Yip, 2002).

**Lexical tone perception**

Like the perception of other linguistic objects, the perception of lexical tone is a complex process that is influenced by characteristics of the speech signal, speaker, and listener. Similar to segmental phonemes, lexical tone is a linguistic category that must accommodate variation from one utterance to another. Even within an utterance, tones of the same type may not have the same F0 or shape. Variance occurs for several reasons, including:

- **Coarticulation** with neighboring tones, resulting in peak delay, or the shifting of F0 targets within the syllable or onto the following syllable (Xu, 1999a, 1999b; Xu, 2001; Xu & Wang, 2001). Acquired knowledge of the acoustic effects of coarticulation may aid in tone identification in context (Gottfried & Suiter, 1997; Xu, 1994).

- **intonation**, requiring listeners to integrate sentence-level prosodic context when identifying tones (Connell, Hogan & Rosspal, 1983; Peng, 1997);

- **Downtrend**, referring generally to any lowering of pitch across an utterance, and including **declination**, a lowering of F0 from beginning to end of an utterance resulting from lowered air pressure in the lungs, and associated phonological processes such as **downrift** and **downstep**, which are language specific grammatical rules resulting in the lowering of tones in certain environments (Connell & Ladd, 1990).

Listeners must also compensate for variation between talkers, a process known as **speaker** or **talker** normalization. If a tone has a distinctive contour, it may be identified in isolation, but in the case of level tones, or of tones with similar contours in different ranges, information about the pitch range used by the talker is necessary to identify the tone, because the pitch range used by speakers can vary considerably based on gender, anatomy, affect, and other factors (Gussenhoven, 2004b; Lee, 2009; Wong & Diehl, 2003). This information can be deduced from external cues, such as the preceding utterance context (Moore & Jongman, 1997; Wong & Diehl, 2003; Xu, 1994;1997), or from internal cues, such as vowel quality and other spectral components like harmonics and voice quality (Honorof & Whalen, 2005; Lee, 2009; Wong, Nusbaum, & Small, 2004).

**Multidimensionality.** Lexical tone is not a unitary phenomenon, but an abstract linguistic object composed of multiple perceptual dimensions, including not only fundamental frequency, but also spectral and temporal components (e.g., duration, amplitude, harmonic rise/fall times). Listeners employ these dimensions to different degrees as perceptual cues to tonal categories based on the relative importance of the dimension in their native tone system (Lee, 2009). These multiple cues provide a degree of redundancy, allowing accurate perception within noise (Kong & Zeng, 2006), or of degraded stimuli, as in whispered speech where F0 information is unavailable (Abramson, 1973; Liang, 1963; Liu & Samuel, 2004).

Pitch is the primary perceptual component of tone, but the pitch component can be further subdivided into multiple perceptual dimensions. Gandour and Harshman (1978) compared the pitch dimensions used by speakers of several languages to identify tones. Thai, Yoruba, and English speakers were asked to rate the similarity of synthesized words that differed based on pitch and length, and contained a variety of level, rising, and falling pitch patterns. Multidimensional scaling analysis revealed five primary dimensions accounting for the perceptual differences between tones:

- **Average pitch**, or **height**, which distinguishes tones based on their average F0 level; thus, it maximally distinguishes high (55) from low (11) tones, and groups rising and falling tones (such as 15 and 51) together with mid-level tones (33), because the average pitch of each is in the middle of the range.

- **Direction**, which distinguishes between rising, falling, and level tones, regardless of their pitch range or degree of pitch change. Thus, it treats a low-rising tone (13), a high-rising tone (35), and a low-to-high rising tone (15) as similar, and distinct from level and falling tones.

- **Length**, which distinguishes words based on duration, a common correlate of tone in many languages.

---

1 Tones are often defined using Chao numbers, which indicate the beginning and ending frequencies of the tone. 1 indicates the lowest frequency, and 5 indicates the highest; thus, 55 indicates a high level tone, and 51 indicates a tone falling from the high end to the low end of the range.
• **Extreme endpoint**, which distinguishes tones which end in the extremes of the pitch range (1 or 5) from tones which end in the middle (3); thus, it groups tones like 11, 15, and 55 together, in opposition to tones like 33 and 53.

• **Slope**, which distinguishes tones based on the steepness of pitch change; thus, it groups tones which change pitch rapidly (15, 51), distinguishing them from tones which do not change at all (11, 33, 55), with less steeply changing tones (35, 53) intermediate between the level and sharply changing tones.

• Gandour and Harshman group the pitch-related dimensions into two categories: static (height, extreme endpoint) and dynamic (direction, slope). They argue that static dimensions reflect general auditory capabilities, whereas dynamic dimensions reflect language-specific perception. This is based on differences observed between Thai, Yoruba, and English speakers, especially the fact that English speakers, who have no experience with lexical tone, demonstrated insensitivity to the dynamic dimensions, relying almost exclusively on static dimensions (especially height). The authors suggest that the English group, as speakers of a non-tone language, were reliant primarily on these non-linguistic, static dimensions, whereas tone language speakers also attended to dynamic properties of pitch (although the possibility that other kinds of prosodic categories influence tone perception by non-tone language speakers has not been entirely ruled out; cf. Francis, Ciocca, Ma & Fenn, 2008).

Subsequent studies by Gandour and others (Avelino, 2003; Chandrasekaran, Gandour & Krishnan, 2007a; Gandour, 1979a; Gandour, 1979b; Gandour, 1983; Khouw & Ciocca, 2007) find a general correspondence between important dimensions across languages using a variety of methodologies, which supports the finding that non-tone language speakers are less sensitive to dynamic pitch than are tone language speakers. For example, Chandrasekaran et al. (2007a) demonstrated that group differences in the amplitude of electrical brain activity (event-related potentials, or ERP) of Mandarin and English listeners in response to Mandarin tones could be predicted by differences in their perceptual weighting of height and contour.

Several studies also indicate that sensitivity to perceptual dimensions of pitch varies not only as a function of the presence or absence of lexical tones in a language, but among individual tone languages, depending on their tone inventories, tonal phonology, and other factors. For example, both Thai and Yoruba speakers attend to direction, but Thai speakers are more sensitive to slope than Yoruba speakers (Gandour & Harshman, 1978). Yoruba, despite its classification as a register-tone language, contains rising and falling tones resulting from phonological rules, thus making direction important. But these tones do not have the same lexical (e.g., meaningful) status as the contour tones of Thai, in which multiple contoured tones must be distinguished from one another on the basis of slope.

Cantonese speakers are more sensitive to tone height than speakers of Mandarin, Taiwanese, and Thai, which is consistent with the Cantonese tonal inventory, in which multiple rising and falling tones must be distinguished on the basis of height, unlike the inventories of the other languages (Gandour, 1983). The importance of height in Cantonese, with its six tones, is somewhat striking, given that this dimension seemed to be the most important, and perhaps only, dimension available to non-tone language speakers in such a task. This illustrates that both static and dynamic dimensions can play a role in native tone language discrimination, and that the shape and configuration of tones in the inventory may be more important than the number of tones. Other differences among these languages, including a greater sensitivity to direction among Thai speakers relative to the Chinese languages (Mandarin, Taiwanese, and Cantonese), were attributed to the presence of phonological rules called tone sandhi, in which certain tonal environments trigger changes in the identity or shape of tones, rather than differences in tonal inventory (Gandour, 1983).

These multiple dimensions of pitch are used redundantly by tone language speakers. Although pitch contour is of primary importance to tone identification in Mandarin, Mandarin speakers can identify the tones of Mandarin words that include only the onset of the word (six glottal pulses, or about 20-30 ms), eliminating much of the dynamic pitch information. This suggests that pitch height (and possibly other non-pitch cues like voice quality) can also be used by Mandarin speakers to identify tones (Gottfried & Suiter, 1997; Lee, 2009).

Conversely, pitch movement also plays a role in the perception of level tones, as vocal pitch is rarely, if ever, completely level, especially in running speech (Abramson, 1978; Abramson, 1975, Gandour & Harshman, 1978). House (1989) found that sensitivity to pitch height and movement is affected by syllabic context, suggesting that the
Research on the relevant perceptual cues for non-Asian languages beyond the study by Gandour and Harshman (1978) is scarce. Connell (2000) compared the ability of English and Mambila (Benue-Congo, Cameroon) speakers to classify synthesized tone stimuli. Like Yoruba, the four level tones of Mambila contain some pitch movement, which plays an important role in their identification. Nonetheless, some Mambila speakers could classify perfectly level synthesized tones, as could English speakers. Differences arose between the groups in how English and Mambila speakers divided the frequency range into categories: Mambila speakers’ division of the frequency range was not uniform; some categories included a greater range of frequency than others. The English-speaking listeners partitioned the pitch space in a more uniform way than Mambila speakers. Importantly, a signal detection analysis revealed no difference in sensitivity to pitch between the Mambila- and English-speaking groups. These results suggest that although the Mambila- and English-speaking groups could perceive the continuous dimension of pitch equally well, the Mambila speakers also used their language-specific knowledge to map pitch onto their tonal categories, while the English speakers’ perception of pitch was driven by other kinds of knowledge, or by ad hoc, stimulus-driven categories.

Avelino (2003) examined the perception of tone continua in Yalálag Zapotec (Otomanguean, Mexico), which has High, Low, and Falling tones. Avelino found that Zapotec listeners use multiple pitch dimensions to discriminate tones, including height, endpoint pitch, and direction, but employ these cues differently for each tone contrast, consistent with findings from other languages.

In summary, tone perception appears to be implemented cross-linguistically by several perceptual dimensions of pitch. These dimensions can be classified as static or dynamic. An important static property is height, and two important dynamic properties are direction and slope of pitch change. The way in which these dimensions are used by speakers in perceiving tones is determined in part by general principles of audition, the tonality (tonal or not) and inventory of tones in the language, and the phonology of the language. Like other aspects of language, as the perceptual system becomes attuned to the tonal properties of one’s native language, the ability to respond to second languages is affected (Pons, Lewkowicz, Soto-Faraco & Sebastián-Gallés, 2009). These effects are reviewed in the next section.

Native and second language tone perception. Findings from second language tone perception are consistent with a general principle of language acquisition: learning affects subsequent learning, focusing the perceptual system on the target language and lowering sensitivity—though not irreversibly—to non-native properties. Evidence of this acquired sensitivity to the perceptual and systematic properties of a native tonal system can be seen in the perception and learning of nonnative tones. The effects of tone language experience are numerous and varied, but can be grouped into three relevant levels. These are

- a pre-linguistic acoustic level, at which acoustic information is extracted and encoded;
- a phonetic level, at which phonetic features are recognized;
- and a phonological level, at which phonetic dimensions are mapped to categorical representations in memory.

These effects will be discussed in reverse order.

Phonological effects. Theories of segmental acquisition and second language (L2) perception (e.g., Best, McRoberts, & Goodell, 2001; Iverson, Kuhl, Akahane-Yamada, & Siebert, 2001), although they differ in details, generally predict that listeners’ native speech categories can both help and hinder perception of new speech sounds by linking them to familiar ones, either facilitating learning by reducing new information, or interfering by making spurious connections which distort the categories of the L2 system.

Some studies suggest that differences between speakers of tone and non-tone languages are qualitative—that tone language speakers process tones in linguistically relevant ways, while non-tone language speakers do not. Gandour (1998) found that Thai speakers show brain activation near Broca’s area (measured with positron emission tomography, PET, which detects metabolic activity) when processing words based on tone, whereas English speakers do not. Broca’s area is associated with processing linguistic structure, suggesting that English speakers are not treating tone as linguistically relevant. Mandarin speakers have a right ear advantage (indicating left hemisphere dominance) for tone processing, whereas English speakers do not (Klein, Zatorre, Milner & Zhao, 2001; Wang, Jongman & Sereno, 2001). This is compatible with traditional models of hemispheric specialization which propose that lexical and
syntactic processing occurs primarily in the left hemisphere, and that prosodic processing occurs primarily in the right. Differences in lateralization suggest that non-tone language speakers do not process pitch information in a lexically relevant way, instead treating it as intonation or non-linguistic pitch. Wang, Sereno, Jongman, and Hirsch (2003a) found that English speakers learning Mandarin shift tone processing to the left hemisphere, indicating that similar lateralization occurs due to second-language tone experience as well.

Another formulation of the qualitative distinction between tone and non-tone languages is the suggestion that speakers of non-tone languages do not perceive pitch in the same categorical fashion as do tone language speakers (cf. Francis, Ciocca, & Ng, 2003, for evidence that tones may not be processed categorically in all contexts by native speakers, either). Hallé, Chang, and Best (2004) showed that French speakers did not perceive Mandarin tones with greater accuracy in between- versus within-category comparisons, a pattern characteristic of categorical perception. The authors do not suggest that French listeners cannot perceive pitch categorically, but that the Mandarin tones are sufficiently different from French intonation categories as to be too difficult to assimilate. Hallé et al. go so far as to say that the French listeners not only fail to perceive the phonemic categories, but that they perceive the tones as nonlinguistic pitch fluctuations.

Peng, Zheng, Gong, Yang, Kong, and Wang (2010) showed that non-tone language (German) speakers could perceive synthesized speech tones categorically, but that their categorical boundaries were less sharp than native tone language speakers. In addition, the location of category boundaries among speakers of different tone languages (Mandarin and Cantonese) seemed to be influenced by the particular tones of their native language (cf. Huang, 2004, for a similar comparison of Mandarin dialects).

Further evidence of the influence of non-tone language speakers’ native pitch knowledge on tone perception was found by Wang, Spence, Jongman, and Sereno (1999), who trained English speakers to perceive Mandarin tones in a lexical identification task. The distinction between Tone 1 (high level) and Tone 4 (falling) was reported to be especially difficult and resistant to improvement. Tones 2 (rising) and 3 (dipping) are also acoustically similar and were initially difficult to discriminate, but English speakers improved on this contrast after training. The authors suggest that this is because Tones 1 and 4 are similar to English stress patterns, and thus more subject to interference and inertia, whereas Tones 2 and 3 are not.

Wang, Jongman, and Sereno (2003b) tracked learners’ perceptual and production performance over time, finding that it was not driven simply by phonological-level categorical changes, but by changes in phonetic sensitivity as well. Pitch contour accuracy improved to a greater degree than did pitch height, supporting the distinction between these phonetic properties found in earlier studies, and suggesting that they are targeted differentially by experience.

Phonetic effects. Wayland and Guion (2003) tapped into this phonological/phonetic distinction, finding that experienced native English-speaking learners of Thai show an effect of interstimulus interval (ISI) in tone discrimination, with better perception of a difficult contrast with shorter ISI, a pattern not shown by native Thai controls. Based on work by Burnham, Francis, Webster, Lukasneyanawin, Lacerda, and Attapaiboon (1996) linking shorter ISI with a phonetic mode of perception and longer ISI with a phonological mode, Wayland and Guion argue that although the experienced English group displays improved phonetic perception compared to naïve English listeners, they lack the ability to use phonemic categories from long-term memory to encode the stimuli phonologically, as native Thai listeners can.

In a subsequent study, Wayland and Guion (2004) examined how this phonetic/phonological distinction changes with experience, training English and Chinese (Mandarin and Taiwanese) speakers to discriminate Thai tones. Before training, Chinese speakers discriminated Thai tones better than the English group but only at short ISI; after training, the Chinese group was better than the English group at both short and long ISI. The Chinese speakers’ advantage before training indicates that speakers of a tone language have an advantage over speakers of non-tone languages in their phonetic mode of pitch processing, but like the non-tone speakers, cannot recruit the phonological categories of the target language. After training, the Chinese group showed an advantage over the English group in both the phonetic and phonological conditions (though still less than native-like), suggesting that with experience, they gained phonological knowledge that the English speakers had not. Tone language speakers seem to have an advantage over non-tone speakers in learning a new tone contrast in two ways: first, they can transfer their
phonetic knowledge (pitch tracking and other cues); second, they can more quickly apply phonological knowledge, either by mapping new tones onto native categories or by learning new categories, an ability that non-tone language speakers lack, because they have not yet acquired the requisite phonetic knowledge.

Francis et al. (2008) describe the explanation presented by Wayland and Guion (2004) as a levels of representation account, contrasting this with the category assimilation account of Hallé et al. (2004), which posits that both tone and non-tone language speakers process foreign tones in relation to their native categories, but that tone language speakers have more success than non-tone speakers because the latters’ native categories are intonational, rather than tonal. The levels-of-representation account supposes that non-tone language speakers have no relevant phonological category to which to relate tones, and therefore cannot use pitch for lexical tasks; the category-assimilation account supposes that non-tone language speakers do use pitch-based categories, but because they are intonational they have less success. It does not suppose any differences in the ability to perceive the acoustic properties relevant to lexical tone.

Francis et al. (2008) also review findings by Wang, Behne, Jongman, and Sereno (2004), who found that Norwegian listeners do not show a left-hemisphere advantage in processing Mandarin tones, as do native Mandarin speakers, even though they are familiar with lexical tone categories in Norwegian, for which they show a lateralization effect. Although Wang et al. conclude that this effect was due to Norwegian listeners’ unfamiliarity with the specific perceptual correlates of Mandarin tones, it is worth noting that in order to equalize error rates across groups, Wang et al. (2004) manipulated ISI and other variables independently for each group based on pilot data; they grouped Norwegian participants together with Mandarin speakers, rather than English controls on ISI, suggesting some advantage in phonetic-level processing by the Norwegian speakers over the English speakers, despite the fact that this does not result in a detectable hemispheric shift for the foreign tones.

Francis et al. (2008) trained Mandarin and English speakers to identify Cantonese tones. Although both groups showed similar overall accuracy on pre- and post-training identification tests, the results of the tests and pre- and post-training difference ratings revealed that initially each group perceived best the tones that could easily be mapped to their native categories. Mandarin listeners placed greater emphasis on direction than on height, confusing tones that shared similar pitch contours, whereas English listeners placed greater emphasis on height than on direction, confusing tones with similar average pitch height. Both groups improved after training and adjusted their weighting of the phonetic properties to more heavily favor height, an especially important dimension of tone in Cantonese (Gandour, 1979a, 1983). The authors conclude that native categories clearly influence how second language categories are initially perceived, and that phonetic features also play a role when new categories cannot be assimilated (or when task demands, such as ISI, preclude access to phonological categories), and perceptual tuning or re-weighting of phonetic properties is a key process during learning.

Neurological evidence for such cue-weighting comes from studies of the mismatch negativity (MMN) ERP component, which functions as a “change detector” within a series of stimuli; greater MMN response indicates greater sensitivity to the item or property that has caused a violation of expectation. Chandrasekaran, Krishnan, and Gandour (2007b) showed that Mandarin-speaking listeners showed a greater MMN response to an acoustically dissimilar set of tones (Tone 1/Tone 3) compared to an acoustically similar set (Tone 2/Tone 3). English listeners did not show such a difference, and Mandarin and English speakers differed in their MMN response only in the high contrast condition. The authors speculate that the English listeners placed greater importance on pitch height, leading them to treat both tone sets as equivalently different, while the Mandarin listeners were influenced by their native language knowledge of the dynamic properties of these tones (i.e., direction and slope), and therefore perceived the tones as less similar to one another. Chandrasekaran et al. (2007a) confirmed this through multidimensional scaling, revealing two dimensions, interpreted as height (linked to average and offset F0) and contour (linked to the rate of F0 change throughout the tone); Mandarin-speaking listeners relied more heavily on the contour dimension than did English-speaking listeners.

Wong and Perrachione (2007) illustrated how sensitivity to these phonetic properties links acoustic perception with linguistic performance. They trained English speakers to associate pseudowords incorporating Mandarin tones with meanings. Learners’ pre-training ability to perceive the relevant pitch contours in a non-lexical (although still
linguistic) pretest predicted their success in learning the lexical task. Performance was also correlated with levels of music training. These findings are interpreted as evidence for a “phonetic–phonological–lexical continuity” in which lower level acoustic and categorical knowledge must be established before being used for lexical tasks. Lee, Perrachione, Dees, and Wong (2007) examined the effect of stimulus variability during such training. Learners with higher pre-training pitch perception ability learned best with a high-variability training set (containing stimuli from four talkers), whereas those with poorer pre-training pitch perception learned best with a low-variability set (containing stimuli from one talker). Accommodating greater variability requires more robust representations, and learners with less initial pitch perception ability have not yet refined the “phonetic categories [which] need to be established before the phonetic details are used phonologically, i.e., to contrast word meanings” (p. 1592). Both groups improved in their ability to perform the lexical task; however those with lower pitch perception ability were relying more on phonetic detail, whereas those with higher pitch perception ability more quickly learned and used categorical (phonological) knowledge. Chandrasekaran, Sampath, and Wong (2010) linked this advantage in tone learning back to phonetic cues to tone, finding that better learners attend more to pitch direction than do poorer learners, and that learning to use Mandarin tones to identify words increases the ability to identify pitch direction.

Acoustic effects. A key question about the results of Wong and Perrachione (2007), Lee et al. (2007), and Chandrasekaran et al. (2010) is to what degree the pre-training differences in pitch perception ability among their subjects are general. Are the observed differences attributable to a task-general pitch perception ability (i.e., at the acoustic level), or only to the ability to use pitch in a speech context (i.e., at the phonetic level)?

Krishnan, Xu, Gandour, and Cariani (2005) found more accurate encoding of pitch in the auditory brainstem by Mandarin speakers compared to English speakers, as measured by the Frequency Following Response (FFR), an EEG-based signal that decodes neural firing patterns into the frequencies they represent. The FFR of Mandarin speakers exhibited stronger representations and smoother tracking of the fundamental frequency and second harmonic of Mandarin tones. Although the differences between the Mandarin and English speakers were specific to linguistically-relevant dimensions, rather than simply to more accurate encoding of all aspects of the signal, these properties (F0 and harmonics) could be relevant to a variety of speech and non-speech auditory tasks. Song, Skoe, Wong, and Kraus (2008) found changes in such F0 encoding by English speakers after only a short period of training on Mandarin tones, but only for Tone 3 (dipping), which was the most difficult tone for English learners before training.

Bent, Bradlow, and Wright (2006) demonstrated that differences in auditory processing between English and Mandarin speakers extend to non-speech sounds. These differences were limited to the discrimination of pitch contours; there were no differences between English and Mandarin speakers on a non-speech frequency discrimination task. The only differences between the groups were on certain falling and flat pitch contours, which Mandarin speakers misidentified more often than did English speakers. Based on a signal detection analysis, the authors suggest that this may be due to response bias on the part of the Mandarin group (i.e., they were treating the pure tones as speech sounds), rather than a difference in sensitivity. Nevertheless, this finding suggests either that perceptual tuning to pitch properties relevant to linguistic representations is available for the processing of pitch information generally, or that non-speech sounds can be processed ‘linguistically’ if they share properties with linguistic representations.

Xu, Gandour, and Francis (2006a) found that the differences between English and Mandarin speakers’ perceptions of level and rising speech tones were mirrored in their perception of similar non-speech sounds, which the authors theorize results from the interaction of domain general sensory memory and long term phonological memory. Importantly, whereas Mandarin speakers perceived both speech and non-speech stimuli in a similar categorical fashion, English speakers perceived the non-speech versions in a more categorical fashion than they perceived the speech sounds. This could indicate that when English speakers perceive pitch in speech, they do not process it categorically because it does not match their stored representations of speech categories. Similarly, Mattock and Burnham (2006) (cf. Mattock, Molnar, Polka & Burnham, 2008) demonstrated that English-learning infants show a decrease in discrimination sensitivity to lexical tones, but not a corresponding decrease in sensitivity to similar non-speech
sounds. This leads the authors to argue for an early separation of speech and non-speech perceptual processes. An alternative explanation is that these results simply illustrate the early emergence of a speech mode of processing (Burnham et al., 1996; Francis & Ciocca, 2003), and the first step in perceptual reorganization for infants is to ignore irrelevant dimensions in linguistic tasks, with loss of sensitivity following later. Indeed, Francis and Ciocca (2003) provide evidence for such a speech mode of processing, finding that Cantonese adult speakers showed an order effect in speech-tone discrimination, but English speakers did not; neither group showed an order effect for corresponding non-speech tones.

Although the results of Xu et al. (2006a) and Mattock and Burnham (2006) might suggest that language experience should not affect the perception of non-speech sounds, further studies demonstrate that the language-induced tuning of pitch representation extends to non-speech sounds that share properties with speech. Xu, Krishnan, and Gandour (2006b) found that English and Mandarin speakers did not differ in their FFR responses to linearly rising or falling non-speech pitches, but they did differ in response to nonlinear changes in non-speech pitch, which more closely resemble Mandarin tones. Chandrasekaran, Krishnan, and Gandour (2007c) extended the results of Chandrasekaran et al. (2007b) and Chandrasekaran et al. (2007a), finding that the MMN response of Mandarin speakers was greater than that of English speakers only for curvilinear pitch contours.

Krishnan, Swaminathan, and Gandour (2009) found that in response to Mandarin tones that were acoustically transformed such that they preserved pitch information while obscuring other speech-specific acoustic information (iterated rippled noise; IRN), the FFR of Mandarin-speaking listeners exhibited smoother F0 tracking and more robust pitch representation. This was manifested not only in more robust representation of fundamental frequency, but in the extended representation of harmonics; the FFR of Mandarin-speaking listeners more accurately encoded pitch information and represented spectral information up to the fifth harmonic, whereas that of English-speaking listeners represented spectral information only up to the third or fourth harmonic. Krishnan et al. (2009) and Xu et al. (2006b) note that the investigation by Xu et al. (2006a) examined only linearly changing pitch contours, and that these linear approximations of the curvilinear Mandarin tones do not adequately capture language-specific tunings resulting from Mandarin tones. These effects are ultimately domain-general, in that the early auditory system shapes multiple acoustic dimensions associated with pitch that are later processed by domain-specific (e.g., language, music) mechanisms.

Summary

Lexical tone is the use of pitch in conjunction with segmental information to convey lexical meaning. Lexical tone is primarily defined by pitch patterns, which consist of several perceptual dimensions found cross-linguistically. Three of the most important dimensions are height, based on the position of the tone within the reference range, direction, based on the direction of change in pitch over the syllable, and contour, based on the rate or shape of pitch changes. The importance of these dimensions to the tone system of a language varies based on its tonal inventory and tonal phonology, and linguistic experience with pitch causes a top-down tuning of pitch perception at phonological, phonetic, and acoustic levels. This tuning, especially at the phonetic and acoustic levels, suggests a route by which linguistic experience may subsequently affect perception in other acoustic tasks, including music. Likewise, if music experience also affects this acoustic level, then such effects are bidirectional, and music experience affects language via a similar route and in similar ways.

Musical tone

Lexical tones are a fundamental unit of pitch in language, and they exist at the phonological level, mediating between acoustic properties of the signal and the meaning of words and morphemes. A note is the smallest “meaningful” unit of music, and notes contain many kinds of information in addition to pitch, including duration, timbre, and loudness (Rasch & Plomp, 1982). Although timbre is important for music, it is generally considered secondary to pitch; in fact, Patel (2008a) argues that this is a key difference between language and music: Linguistic categories are primarily timbre-based, but musical categories are primarily pitch-based. Lexical tone is an exception to the first generalization, but it remains that every language contains timbral categories (e.g., consonants and vowels), while not every language has lexical tone.

I refer to that part of a note based on pitch as musical tone, just as the component of a word’s meaning determined by pitch is a lexical tone. Like a lexical tone, a musical tone
is an entity in a rule-governed system, and can be given various names depending on its function with the system, such as B-flat, or do. This section will explain some acoustic facts related to musical tones, how such tones and their relationships are perceived, and how musical objects like melodies are built from individual notes, with particular emphasis on characteristics of musical tone perception that parallel those found in language.

**Acoustic and musical facts**

Burns and Ward (1982) review the organization of frequency into musically meaningful units. Pitch categories in music consist of a finite set of tones, whose relationships to one another are defined by the ratio of their frequencies, or interval. Modern Western\(^2\) music divides the octave into 12 equal intervals, with a subset of the 12 making up the scale on which a given composition is based (12-tone serialism is a notable exception). This scale forms the basis for the melodic and harmonic structure of the music, and this melodic structure is an important link between music and language.

A melody is a collection of individual notes arranged linearly. The temporal dimension, or rhythm, of a melody is also an important component, but will not be discussed here. Apart from its implied harmonic structure and rhythm, the content of a melody can be described in at least three ways: by its key, contour, and intervals. Contour encodes the up/down pattern of pitch changes, regardless of the note identities; interval encodes the precise distance from each note to the subsequent one. Key is a complex term, but includes the hierarchical relationship among notes in a melody, indicating the tonal center and underlying scale. In most cases, identical melodies that differ only in their tonic or starting note are considered to have the same musical identity.\(^3\) Experimental evidence suggests that these are dissociable perceptual dimensions, with unique developmental courses (Trehub, Bull & Thorpe, 1984; Trainor & Trehub, 1992; Trehub, Schellenberg & Kamenetsky, 1999). Under most conditions, the contour of a melody is remembered more accurately by listeners than its interval content (Bartlett & Dowling, 1980; Edworthy, 1985; Peretz & Babaï, 1992; Trainor, Dejardins & Rockel, 1999), and sensitivity to interval and contour is enhanced by training (Fujikata, Trainor, Ross, Kakigi & Pantev, 2004).

**Musical typology.** Most of the descriptions, generalizations, and experimental studies of music and musicians discussed here pertain to the Western (European) musical tradition. Although this tradition is better studied from a theoretical and cognitive perspective, it is important to consider how the properties of other musical systems differ from or resemble Western music. Due to increasing globalization, Western music is pervasive in cultures around the world, but may interact in complex ways with the native musical systems of listeners from different cultures (Demorest, Morrison, Beken, & Jungbluth, 2008), making cross-cultural studies particularly difficult. The music of the various cultures of the world includes a great diversity of scales and harmonic systems besides that found in the Western tradition, such as Indonesian gamelan (Vetter, 1989) and Indian ragas (Castellano, Bharucha & Krumhansl, 1984). In addition, some styles of contemporary Western art music have developed tonal and atonal systems quite distinct from the roots of European folk music, such as 12-tone serialism (see Peretz, 2006 for an interesting discussion of the psychological validity of such musical idioms).

Krumhansl (1990) reviews the results of several studies on the perception of tonality in Western and several non-Western musical forms, including 12-tone serial, North Indian, and Balinese (gamelan) music. Listeners from each of these cultures (Western musicians well-versed in atonal music in the first case) were found to perceive the tonality and scale structure of the music in a manner consistent with music-theoretic descriptions of the relevant scale structure. Even non-enculturated listeners (e.g., Westerners listening to gamelan music) could extract some information about pitch relations from the regularities in the non-Western music; however, they did not perceive pitch relationships that depend on knowledge of the underlying scale structure.

Music is a culturally acquired system in a manner similar to language—according to innate predispositions that guide and shape the extraction of culturally specific patterns from the environment (Lynch & Eilers, 1991; Trehub et al., 1999; Trehub, 2003a; Trehub, 2003b). Universal properties of musical systems may allow even non-native listeners to extract some structure from an unfamiliar musical idiom.

Lynch and Eilers (1991) found that neither musician nor nonmusician Western children (10–13 years old) could

---

\(^2\) See the next section for a discussion of how other musical systems may differ from the descriptions given here.
detect mistunings in the Javanese gamelan pelog scale, although both groups could detect mistunings in Western major and minor scales, with musicians outperforming nonmusicians in this task. This study demonstrates the effects of training at an early age, but another important finding is that even nonmusician children demonstrate knowledge of the scales of their native musical forms. This task draws less on musical ability than on knowledge of musical structure, and indicates that even trained musicians cannot necessarily generalize this culture-specific knowledge to another musical context.

Can the particular musical system acquired affect domain-general auditory perception? The available evidence suggests that sensory tuning in response to music drives auditory changes similar to the effects of native language on the perceptual system. Like language, the stored representations of musical units change how sound is processed. Cooke (1992) examined the tuning tolerances of Ganda and Soga (Uganda) musicians, who use an equi-pentatonic scale (five notes relatively equally spaced along the octave). The tolerances of the Ugandan musicians were larger than those of the Westerners, an effect that was linked to the larger interval size of the pentatonic Ugandan system compared to the smaller intervals of the Western 12-note system; a pitch in a pentatonic scale can tolerate more variation before it intrudes on a neighboring pitch category, and so the categories are wider.

Schneider (2001) notes the difficulties inherent in describing pitch systems, which result from the fact that the percept of musical pitch is more than simply fundamental frequency, although this is the only pitch-related acoustic property that Western music theory encodes. Many kinds of music (e.g., gamelan) rely on instruments (gongs and metallophones), which produce sounds with inharmonic spectra and weak fundamentals, leading to ambiguity of pitch. These seem to be important features of gamelan music, which relies on interactions of these complex harmonics between instruments to create a shimmering perceptual effect, and may account for the wide variations in tuning found between different ensembles. Schneider (2001) speculates that “these components constantly interfere with the base [fundamental] frequency components and may be heard as a sequence with a melodic line of its own” (p. 513). It appears that in music, just as in language, there are multiple cues used to perceive the basic elements (notes) of the system. The relative importance of these cues depends on the musical idiom, and the perceptions of listeners of a particular type of music are influenced by its use of these cues.

Musicality

Another way of considering the effect of music on the perceptual system is the degree of musical competence, or musicality, possessed by the perceiver. Formal training includes not only practice in producing music, but also the development of an ability to perceive fine auditory distinctions. These highly developed motor and perceptual skills appear to drive neuroplastic changes in the brains and behaviors of musicians (Besson, Chobert & Marie, 2011a; Münte, Altenmüller & Jäncke, 2002).

Training has been shown to affect higher cognitive functions such as auditory attention (Strait, Kraus, Abecassis & Ashley, 2010a; Strait, Kraus, Parbery-Clark & Ashley, 2010b). Musicians demonstrate superior temporal processing on several kinds of tasks (Rammsayer & Altenmüller, 2006), and their ERP responses indicate greater sensitivity to changes in the temporal structure of music, even preattentively (Rüsseler, Altenmüller, Nager, Kohlmetz & Münte, 2001; Tervaniemi, Ilvonen, Karma, Alho & Näätänen, 1997). The brains of musicians show structural differences compared to nonmusicians, including more connections at the corpus callosum, which connects the two hemispheres (Schlaug, Jäncke, Huang, Staiger & Steinmetz, 1995a; Schmithorst & Wilke, 2002) and increased gray matter in areas associated with auditory functions, and with motor, visual, and higher cognitive functions as well (Gaser & Schlaug, 2003).

Functional Magnetic Resonance Imaging (fMRI) studies indicate that musicians activate more cortex in response to piano (Pantev, Oostenveld, Engelien, Ross, Roberts & Hoke, 1998) and sinusoidal tones (Schneider, Scherg, Dosch, Specht, Gutschalk & Rupp, 2002) and their early cortical ERPs to pure, piano, and violin tones display a larger amplitude than do those of nonmusicians, correlating with degree of training (Shahin, Bosnyak, Trainor & Roberts, 2003). Shahin, Roberts, and Trainor (2004) found similar results even for four- and five-year-old children enrolled in music lessons. Pantev, Roberts, Schulz, Engelien and Ross (2001) found that musicians’ neural responses were heightened for the timbre of their own instruments; that is, trumpeters responded more strongly to trumpet sounds than to violin sounds, whereas violinists showed the
opposite pattern. Even short-term training in melodic tasks results in plasticity and an expanded representation of scale tone in the auditory cortex (Pantev, Ross, Fujioka, Trainor, Schulte & Schulz, 2003).

Despite this body of evidence, musicality/musicianship remains difficult to consistently operationalize in research, both because of the various criteria used to define musician by different researchers, and because of cultural differences in musical practice and professionalism. It is also difficult or infeasible to conduct truly randomized studies of the effects of long-term training; thus, it is difficult to rule out the possibility that the differences between musicians and nonmusicians enumerated above result from pre-existing differences between self-selecting groups. The fact that these effects correlated with length of training and age of training onset suggest—but do not definitively establish—a causal link, and the evidence remains somewhat circumstantial. Although not denying a role for inheritance, Sloboda (1994) argues that becoming a skilled musician is more a matter of practice and experience than of innate ability or predisposition, and that the basic perceptual skills necessary for music are possessed in some form by nearly everyone as members of a musical culture. Musicians develop these to a higher level. Although Smith (1997) reviews studies illustrating differences between musicians and nonmusicians, he argues for the inclusion of nonmusicians in music cognition studies, because music is an important part of the cultural and cognitive makeup, even for the unskilled listeners, and a theory of music ought to describe this.

Of course, nonmusicians are not musically naïve; even non-performers grow up immersed in their musical culture, and become familiar with the structure of music. Bigand and Poulin-Charronnat (2006) review studies illustrating musical perception skills displayed by nonmusicians, who the authors describe as experienced listeners, despite a lack of explicit training. Nonmusicians behave similarly to musicians in their experience of tension and relaxation in music, anticipation of musico-syntactic events, and emotional content. Bigand and Poulin-Charronnat note that many experimentally demonstrated differences between musicians and nonmusicians are differences in the magnitude, rather than in the kind, of response. Bigand (2003) argues that experienced listeners behave very similarly to musicians in regard to perception of musical structure, but that they may diverge in their sensory processing of musical tone, although these differences may in some cases be simply a matter of degree.

As noted above, one of the difficulties in evaluating findings from studies that examine the effects of music training is that researchers may use different criteria to qualify participants as musicians. The musicians in studies by Alexander, Wong, and Bradlow (2005), Wong and Perrachione (2007), and Song et al. (2008) were those who had received 6 or more years of vocal or instrumental training beginning before age 10, whereas those in other studies (e.g., Schlaug, Jäncke, Huang, & Steinmetz, 1995b) were defined as active professional musicians. Clearly, musicianship is a gradient ability, from the casual music fan to the performing professional.

The majority of music perception experiments examine musicians alone or in comparison to nonmusicians. Music training affects musicians’ perception of the sounds of music in several ways, including pitch, timbre, and timing (Kraus, Skoe, Parbery-Clark & Ashley, 2009). The effects of music experience on note perception will be reviewed below, with particular attention devoted to sensory and structural properties of musical tone and melody, analogous to the phonological, phonetic, and acoustic levels described for lexical tone.

**Phonological effects.** In the phonological vein, Tervaniemi, Rytikönen, Schröger, Ilmoniemi, and Näätänen (2001) found that the MMN (as measured electrically with ERP, and magnetically, with magnetoencephalography or MEG) of musicians was sensitive to changes in the contour of a musical figure presented at different frequencies, reflecting invariance and generalization across keys. This is analogous to speaker normalization performed during linguistic processing. Subjects who could not accurately discriminate the deviant pattern, whether musicians or nonmusicians, did not show the effect, whereas the effect for those who could discriminate it grew over the course of the experience, reflecting short-term learning. The musicians in the study who most often displayed this effect were those whose experience was in genres that often do not use a score (e.g., pop, jazz), rather than classical music, in which musicians typically perform from music notation; thus, different kinds of musicality may rely more heavily on various modes of learning, and therefore shape perception differently.

Ashley (2003) reported on an experiment designed to illuminate characteristics of musicians’ representations of
Pitch perception

pitch space. He notes that many theorists and researchers have used spatial or geometric representations for pitch, such as the spiral scheme including pitch height and chroma described by Krumhansl (1990). Subjects heard 33 randomized notes distributed over a 7-octave range in each of three synthetic instrument timbres. Subjects completed categorical and continuous pitch height judgment tasks for each note, and the results indicated that musicians seem to use ordinal, linear, and metric representation of pitch height in the task.

Phonetic and acoustic effects. Musicians display a greater MMN response to deviations in both melodic contour and interval structure compared to nonmusicians, but these groups show no difference in their response to frequency changes of pure tones (Fujioka et al., 2004; Pantev et al., 2003). This indicates that the differences resulting from music experience are related to dynamic dimensions of musical pitch, rather than finer pitch discrimination.

Micheyl, Delhommeeau, Perrot, and Oxenham (2006) examined the pitch discrimination ability of professional musicians and nonmusicians, finding a much smaller discrimination threshold for musicians, although nonmusicians could reach musician-like levels with only a few hours of training. The musicians’ advantage held for both pure and harmonically complex tones, but was more pronounced for the complex stimuli. Musacchia, Sams, Skoe, and Kraus (2007) demonstrated more robust FFR representation of F0 and earlier onset of response to musical stimuli by musicians compared to nonmusicians. F0 encoding was correlated with length of training, which the authors argue shows that the observed effects are due to training, rather than innate predispositions among the groups.

Bidelman and Krishnan (2009) examined interval perception using the FFR, finding that responses to dichotically-presented consonant intervals were more robust than were those to dissonant intervals. The authors “infer that the basic pitch relationships governing music may be rooted in low-level sensory processing and that an encoding scheme that favors consonant pitch relationships may be one reason why such intervals are preferred behaviorally” (p. 13165; emphasis added). Such low-level architecture might be recruited by linguistic as well as musical tone perception.

Lee, Skoe, Kraus, and Ashley (2009) examined the brainstem encoding of harmonic (simultaneous) music intervals by musicians and nonmusicians. For both consonant (major sixth) and dissonant (minor seventh) intervals, the FFR of musicians represented the harmonics (especially the second harmonic) of the upper tone of the interval more robustly than did the FFR of nonmusicians. According to the authors, the specificity of this effect for the upper tone reflects its compositional relevance and prominence in previous psychophysical and neurological studies; likewise, harmonics are particularly relevant for the detection of consonance and dissonance, and for the perception of timbre. The effect upon the second harmonic above all others parallels the findings of Krishnan et al. (2005) for lexical tone perception.

Some studies note an effect of music training only on neural mechanisms for perception of abstract musical information, to the exclusion of basic sensory information (Fujioka et al., 2004; Pantev et al., 1998). Others suggest that the effect of musicality may extend to more basic sensory processing (Lee et al., 2009; Musacchia et al., 2007; Schneider et al., 2002), at least for acoustic properties that are musically relevant.

Differences between musicians and nonmusicians in behavioral performance, brain structure, or neural response are not alone sufficient to conclude that music training caused these changes, and are not sufficient to reject the alternative hypothesis that those with preexisting advantages in pitch or rhythmic perception are more likely to pursue music training. In many studies, however, not only is there a group difference between musicians and nonmusicians, but these effects correlate with degree, length, or type of training (Musacchia et al., 2007; Pantev et al., 1998; Shahin et al., 2003; Shahin et al., 2004; Tervaniemi et al., 2001), which is typically interpreted to suggest that the effect is due at least partially to experience.

Summary

Taken together, the results of this body of research indicate that the human auditory system is tuned to perceive acoustic properties relevant to musical structure, even among listeners who are not formally trained musicians. Some of these properties (e.g., F0 height, pitch change over time, spectral composition) are relevant to both language and music, suggesting a possible explanation of the effect of musicianship on performance in second-language acquisition (Wong & Perrachione, 2007). Thus, it appears that in some ways musicians resemble speakers of tone
languages in the way they perceive pitch information. What is not yet known is whether the reverse effect holds; that is, whether tone language speakers resemble musically experienced listeners in the ways in which they perceive musically-relevant aspects of pitch.

**Connecting pitch in music and language**

Kraus and Banai (2007) review research on how auditory processing changes in response to environmental and learning experiences. Many of their findings suggest that auditory processing malleability is controlled in large part by top-down (higher cognitive) functions. They review findings on the loss of sensitivity to non-native phonemes in infants, and enhancement of pitch processing and subcortical encoding among speakers of tone languages. Musicians’ brains also respond more strongly to sound, and show the same subcortical enhancement as do tone-language speakers.

Changes in the auditory environment during development influence the structure of the auditory cortex as well, and cortical reorganization due to sensory loss is not uniform; it shows asymmetries particularly suited to regaining function (e.g., visual cortex is used for hearing in the periphery by individuals with visual impairments). Cases in which learning transfers from perception of one stimulus to perception of another are particularly interesting; perceptual learning can take place as a result of several kinds of neural changes (e.g., increased amplitude, increased precision, sharpened receptive fields, or reorganization of cortical maps). What is clear is that more learning takes place when the relevant characteristics of the stimulus are actively learned and attended to, compared to simple exposure. Two cognitive theories about language and music, the Shared Sound Category Learning Mechanism Hypothesis (SSCLMH; Patel, 2008b), and the OPERA hypothesis (Patel, 2011, 2012), in connection with a more general neurobiological model of learning, Reverse Hierarchy Theory (RHT; Ahissar & Hochstein, 2004; Ahissar, Nahum, Nelken & Hochstein, 2009), provide a promising framework for uniting many of these observations.

Kraus and Banai (2007) cite RHT to account for the influence of top-down processes on perceptual learning, and one that might also explain transfer of learning between domains such as language and music. RHT claims that neural changes begin at the highest cognitive level that can solve a task, with changes to lower areas following when needed in order to provide more accurate sensory input to the task. The tuning of cortical and subcortical pitch representation resulting from lexical training (Song et al., 2008; Wang et al., 2003a) fits within this model, as do similar findings for vision demonstrating lower-level sensory tuning following visual categorization training (Jiang, Bradley, Rini, Zeffiro, Vanmeter, & Riesenhuber, 2007). Due to the presence of both feedforward and feedback connections in the human perceptual system, plasticity occurs at all levels.

RHT could link the parallels between language and music discussed by Patel (2008b) and others, which are united by certain abstract features; perceptual learning related to these features has the potential to cross domains, and even push down into lower level perceptions of both language and music. RHT is compatible with the Shared Sound Category Learning Mechanism Hypothesis (SSCLMH) articulated by Patel (2008b), which asserts that, although language and music have very different structures, they share cognitive and neurobiological mechanisms of sound category learning; that is, language and music are two different systems of sound categories acquired by the same (or some of the same) perceptual processes. The SSCLMH makes claims that contradict some other models of language and music processing (e.g., Peretz & Coltheart, 2003) that suppose the modularity of music and language extends to lower level auditory processing; however, they are in accordance with recent findings by Merrill et al. (2012), which trace the hierarchy of pitch representation in the brain for song and speech, finding a great deal of overlap for linguistic and musical representation in temporal cortical areas, with divergent areas speech and music representation in frontal and parietal regions. Brown, Martinez, and Parsons (2006) made a similar finding for brain areas involved in musical and linguistic production. This fits the predictions of RHT and SSCLMH, in that more overlap is found in temporal resource networks encoding basic properties of sensory stimuli, and less overlap is found among regions higher in the hierarchy, which encode more abstract, task-specific representations. This contradicts other views of acoustic processing (e.g., Zatorre, Belin, & Penhune, 2002) which claim that music and speech diverge earlier in the processing stream, but is consistent with developmental theories linking music and language learning to general perceptual processes and learning mechanisms (McMullen & Saffran, 2010; Trehub & Hannon, 2006).
The framework provided by RHT predicts that cross-domain effects should be found in both directions for properties that music and linguistic tone have in common. The OPERA hypothesis (Patel, 2011, 2012) articulates learning conditions relevant to plasticity. Crossover of music experience to language perception occurs because of (a) the overlap between neural resource networks for language and music, (b) the precision of representations required by musical categories, (c) the emotional connections created by music, (d) the repetitive nature of music training, and (e) the attention it requires. Together, these conditions predict that music experience is more likely to affect linguistic perception than the reverse case. Recent findings demonstrating the influence of linguistic experience on music perception and vice versa can be interpreted in light of these models.

Effects of Music and Musicality on Language Perception

Music training or musicianship has been observed to affect phonological processing and early reading ability (Anvari, Trainor, Woodside, & Levy, 2002), foreign language pronunciation skills (Milovanov, Pietilä, Tervaniemi, & Esquef, 2010), intonation analysis skills (Dankovičová, House, Crooks, & Jones, 2007), perception of speech in noise (Parbery-Clark, Skoe, & Kraus, 2009a; Parbery-Clark, Skoe, Lam, & Kraus, 2009b), lexical stress processing (Kolinsky, Cuvelier, Goetry, Peretz, & Morais, 2009), enhanced reading and speech pitch perception (Magne, Schönh, & Besson, 2006; Marques, Moreno, Luis Castro, & Besson, 2007; Moreno, Marques, Santos, Santos, Castro, & Besson, 2009; see also Schellenberg, 2003, for a review of other “side effects” of music). Some links have also been made between language processing disorders and music perception (Jentschke, Koelsch, Sallat, & Friederici, 2008). Patel and Iversen (2007) propose that at least some of these effects can be explained by the tuning of auditory cortex as a result of music training. Supporting evidence comes from studies of music and tone perception.

Effects of music and musicality on tone perception. Alexander et al. (2005) compared the tone identification and discrimination abilities of American English-speaking musicians and nonmusicians, none of whom had any prior exposure to the Mandarin tones tested. The musicians were both faster and more accurate than the nonmusicians at identifying and discriminating syllables differing only in tone (though still not as accurate as native Mandarin speakers). This suggests that the musicians could transfer some part of their musical knowledge, which the authors suggest is pitch processing ability, to the linguistic task.

Delogu, Lampis, and Olivetti Belardinelli (2006; 2010) tested non-tone language speakers on a memory task requiring them to detect changes in the segmental or tonal content of a list of Mandarin syllables. All subjects were better at detecting segmental changes than tonal changes, but the subjects who demonstrated better melodic memory on a musical intelligence test showed better performance than did those with poorer melodic memory, but only for lexical tone detection, showing that music experience improves linguistic pitch perception, but not other phonological abilities. Further research by this group (Marie, Delogu, Lampis, Olivetti Belardinelli, & Besson, 2010) indicates that music experience could affect non-pitch (segmental) aspects of foreign language perception as well. This could also be driven by non-pitch aspects of music (Besson, Chobert, & Marie, 2011b; Kraus et al., 2009; Kraus & Chandrasekaran, 2010).

Mandarin speakers showed greater MMN response than did English-speakers in response to non-speech homologues of Mandarin tones, both for within- and between-category contrasts (Chandrasekaran, Krishnan, & Gandour, 2009). English-speaking musicians showed some similarities to native Mandarin speakers, demonstrating larger MMN responses to all contrasts than were observed in nonmusicians, though not as large as those observed in native Mandarin speakers. Because the English-speaking musicians had no previous exposure to Mandarin, their larger MMN responses cannot be attributed to knowledge of Mandarin phonetic categories; the authors conclude that it must arise from domain-general enhancement of F0 encoding at the early cortical or subcortical levels (Krishnan et al., 2005; Wong, Skoe, Russo, Dees, & Kraus, 2007). Mandarin speakers actually displayed less accurate behavioral discrimination of the within-category contrast than did English speakers, despite their more robust MMN response.

Native speakers treat these exemplars as belonging to the same category, whereas naïve listeners do not, meaning that this difference must arise due to the language-specific phonological knowledge of the Mandarin speakers. Although the early cortical responses of musicians were similar to those of native Mandarin speakers, they were not identical, indicating that the effects of native language experience and music training may not be qualitatively or...
quantitatively the same; Wong et al. (2007) speculate that learning behaviorally relevant pitch patterns (e.g., music for musicians, lexical tones for tone-language speakers) could transfer to behaviorally irrelevant patterns (e.g., lexical tones for English-speaking musicians) via short-term plasticity, whereas responses to behaviorally relevant stimuli are enhanced by long-term plasticity in the form of cortical and subcortical neural reorganization. Thus, their explanation supposes the quantitative differences in MMN responses observed may be due to qualitatively different processes of perceptual learning.

Musacchia et al. (2007) also demonstrated more robust FFR representation of speech F0 by musicians (no language background was reported) for speech syllables that were not based on Mandarin or another tone language; musicians also demonstrated earlier onset of FFR responses. Strength of F0 encoding by musicians was correlated with length of training and amount of exposure to music. Wong et al. (2007) demonstrated that such transfer may take place specifically as a result of changes in F0 tracking in the auditory brainstem. Recall that Krishnan et al. (2005) demonstrated more accurate encoding of pitch in the auditory brainstem by Mandarin speakers than by English speakers, and that Song et al. (2008) demonstrated similar changes in F0 encoding as English speakers learned Mandarin tones. Wong et al. (2007) compared the FFR in response to Mandarin tones of English speakers who had different levels of music training. In addition to faster and more accurate perception of the tones, the FFR of musicians illustrated a more faithful neural representation of F0 contour; in particular, the accuracy of encoding of Tone 3 (the most difficult tone for English listeners) correlated significantly with perception of the tone and with years of music training. Chandrasekaran, Kraus, and Wong (2012) identified the inferior colliculus, which connects brainstem areas involved in audition with the auditory corex, as an important brain region for such pitch representation and for success at foreign language tone perception. Further support for the idea that the dynamic representation of pitch is specifically enhanced by music experience comes from Wayland, Herrera, and Kaan (2010), who found that musicians were better than nonmusicians at lexical tone contour categorization, but that the groups did not differ in simple frequency discrimination; both groups improved similarly in lexical tone categorization after training.

**Effects of Language Experience on Music Perception**

Experiences with language have been shown to affect music perception. Most of these effects have been observed in comparisons of tone and non-tone language speakers, but some, including linguistic differences in responses to the tritone paradox, have been demonstrated in speakers of both tone and non-tone languages, as well as in speakers of different non-tone languages (Deutsch, 1991; Deutsch, Henthorn, & Dolson, 2004; Dolson, 1994). In the tritone illusion, complex tones with no clear central pitch (Shepard tones) are used to create note pairs that are ambiguous between a rising and falling interval—some listeners hear them as rising, and some as falling, but correlations have been found between linguistic background of the listener and the direction the interval is perceived. Deutsch and colleagues argue that this effect arises from the pitch patterns, including intonation, in the speech of the listener’s community.

**Absolute pitch.** An important language-to-music transfer effect concerns the prevalence of absolute pitch among different linguistic populations. Absolute pitch (AP) is the ability to identify the name of a specific note without a reference pitch. Although it was originally thought only to occur among musicians, Levitin (1994) found that when nonmusicians are asked to sing the melodies of their favorite popular songs, their production is very close to the pitches of the original recordings. Levitin makes a distinction between pitch memory, which does not necessarily require any music education of knowledge, and pitch labeling, which requires a specialized skill. Bergeson and Trehub (2002) made a similar finding for pitch consistency in mother song. Among musicians with pitch-labeling AP, Schlaug et al. (1995b) found that musicians evince greater left-lateralization of activity in the planum temporale, also associated with language (Wernicke’s area), than do non-AP musicians and nonmusicians.

A greater prevalence of AP has been observed among Mandarin-speaking musicians than among English-speaking musicians (Deutsch, Henthorn, Marvin, & Xu, 2006). The cause of this is unclear, because lexical tones are not defined by absolute pitch values. Because AP is defined by the ability to assign labels to pitch categories, the prevalence of AP among Mandarin speakers could be due to domain-general improvements in the ability to assign labels to pitch categories (Lee & Hung, 2008; Lee & Lee, 2010) or to
Absolute pitch appears to be a complex ability that may be influenced by cultural, linguistic, genetic, and musical factors (Hove, Sutherland & Krumhansl, 2010; Levitin & Rogers, 2005). However, because this ability is far from universal, even among musicians, it may be more relevant to consider the effects of language on relative pitch.

Ethnicity. Hove et al. (2010), inspired by findings regarding absolute pitch, conducted two studies designed to illustrate the effects of ethnicity, rather than tone language, on music perception. In the first, they found that Chinese nonmusicians identified melodic intervals more accurately than did both English and Hmong (a contour tone language of East Asia) nonmusicians, who did not differ from each other. In the second, Chinese and Korean (a non-tone language) nonmusicians performed better than English speakers in a similar task. Hove et al. interpreted these results as evidence for an effect of ethnicity (broadly defined), independent of tone-language experience, because the Hmong (tone) group did not differ from English speakers, whereas the Korean (non-tone) group did. There are several difficulties in interpreting these findings.

First, the tasks used by Hove et al. (2010) were labeling and identification tasks, which may tap into abilities other than recognition and discrimination (Levitin, 1994), and which may be affected differently by language and other factors. Second, the assumptions about the effects of each language are not made entirely clear. Although Mandarin and Hmong are both contour tone languages, they are very different typologically and phonetically. The tone inventory of Hmong is complex, with phonation type and spectral information playing important roles in perception (Garellek, Esposito, Keating, & Kreiman, 2012). It is not clear what predictions these phonetic elements make about interval perception, if any. In addition, the only intervals tested were isolated or repeated rising intervals. Mandarin includes more than one tone with a rising component (Tones 2 and 3); Hmong (depending on dialect) does not.

Given the specific nature of many crosslinguistic differences in sensitivity to pitch information, it would not be surprising if Mandarin speakers showed particular sensitivity to rising lexical tones and melodic intervals due to the importance of rising pitch in their language. Likewise, due to the limited stimuli tested, it is difficult to say why the Korean speakers patterned with the Mandarin speakers, given that speakers of different non-tone languages can...
display differences in pitch perception based on the prosodic patterns of their language (Deutsch, 1991). Nonetheless, cultural, educational, and genetic (Dediu, 2010; Dediu & Ladd, 2007) factors may indeed play some role in pitch perception, but the role of linguistic experience must be accounted for through carefully controlled investigation before the magnitude of these other factors can be determined.

Language tonality and melody perception. As discussed earlier, it appears that musicians who do not speak a tone language resemble tone-language speakers in some of the ways they perceive pitch; that is, music experience improves the perception of dynamic properties of pitch in language. We know less about how tone language speakers resemble musicians (that is, whether language experience enhances pitch perception in music), although a few studies have investigated this issue.

Stevens, Keller, and Tyler (2004) found that Thai speakers were faster (but not more accurate) than English speakers at detecting changes in two-note melody pairs. Alexander, Bradlow, Ashley, and Wong (2011) found that Mandarin nonmusicians discriminated five-note melodies more accurately than did English nonmusicians, but English listeners identified (matched with graphical representations) the melodies better than the Mandarin listeners did. Pfordresher and Brown (2009) found that tone language-speaking nonmusicians (Mandarin, Cantonese, and Vietnamese) more accurately discriminated and imitated two-note melodies than did English-speaking nonmusicians, but there was no difference in absolute note errors. Likewise, the tone language group more accurately discriminated intervals but not individual notes compared to the non-tone group. In general, these findings are consistent with a general hypothesis that the effects of language tonality on melody perception are narrow. More studies in this area are needed to make any of these claims more specific.

Remaining Questions and New Hypotheses
The perception of both lexical tone and musical melody relies on dynamic pitch information. Music experience influences the perception of linguistic pitch, and speakers of tone languages perceive music differently than do speakers of other languages, but these effects have not yet been fully explained. Among the research questions asked by Kraus and Banai (2007) is “which acoustic elements of sounds are critical for language and music” (p. 109) in ways relevant to their mutual influence. Because tone language speakers are more sensitive than non-tone language speakers to dynamic aspects of pitch in speech, effects of tone language experience on music perception are likely to be found in aspects of music that rely on dynamic pitch information.

Two properties of musical melody offer a potential homologue to the dynamic pitch dimensions important for tone perception. Contour encodes the up/down pattern of pitch changes in melody, regardless of the note values; interval encodes the distance from each note to the subsequent one. Because melodic contour encodes the direction of movement from one note to the next, it corresponds to the linguistic dynamic tonal dimension direction. Interval, which describes the magnitude of change from one note to the next, corresponds to the dynamic tonal dimension slope. Musical key encodes several properties, such as the scale type (e.g., major, minor) and tonic note (i.e., scale name; A, B, etc.). The pitch levels used in lexical tone do not have the same hierarchical structure as musical scales, but height and tonic are both relatively static properties, representing the location of the melody or tone within the range of audible pitches.

If corresponding properties rely on similar domain-general auditory mechanisms, then RHT (Ahissar et al., 2009) predicts that crossover effects of linguistic or music learning should be limited to dimensions corresponding to those that are tuned, and should not affect other dimensions. For example, a tone language that places great emphasis on direction would affect how its speakers perceive melodic contour, independent of music training; likewise, music training emphasizing intervals would result in changes in sensitivity to the slope of linguistic pitch. This general hypothesis, that the effects of pitch experience on tone and melody perception are not across-the-board but are limited in scope, is consistent with many of the findings already discussed, such as those suggesting that the effects of musicianship on tone perception are driven by more accurate encoding of dynamic pitch information (Wong et al., 2007).

Effects in the other direction (language to melody) have been explored less often, and are somewhat more difficult to interpret within this framework, perhaps because as the OPERA hypothesis (Patel, 2001, 2012) predicts, such effects may be weaker, on average, than those from music to language. Stevens et al. (2004) drew an explicit parallel
between contour in music and language, finding that Thai
speakers were faster than English speakers at detecting
changes in the contour and intervals of two-note melodies
(transpositions and tonic changes were not tested). The
authors claim that this indicates a specific effect, rather than
general improvement of pitch perception. However, because
no differences were found between different kinds of
melodic changes (contour vs. interval), this result does not
rule out a general pitch attunement. Alexander et al. (2011)
showed that Mandarin nonmusicians discriminate melodies
more accurately than do English nonmusicians, but the
types of melodies used did not separately address contour,
interval, and tonic changes. Pfordresher and Brown (2009)
found that tone-language-speaking nonmusicians
discriminate melodies and intervals better than do English-
speaking nonmusicians. The fact that no differences in
discrimination of individual musical tones were found
supports the idea that the language effects are not a general
attunement to pitch; however, the relationship between
contour and interval was not fully explicated in this study
because of the short (two-note) stimuli used.

Bradley (2012) explicitly investigated the effects of
language on contour, interval, and tonic/key perception
using a modified version of the Musical Ear Test (Wallentin,
Nielsen, Friis-Oliverius, Vuust, and & Vuust, 2010) to
examine Mandarin- and English-speaking nonmusicians’
sensitivity to changes in the contour, intervals, and tonic
note of melodies. Mandarin speakers outperformed English
speakers in melodic discrimination overall (consistent with
previously demonstrated advantages for Mandarin speakers),
but this effect held only for contour and interval conditions,
and not for tonic (for which both language groups
performed well above chance). This is consistent with the
importance of the dynamic properties direction and slope in
the tone inventory of Mandarin. Comparing Yoruba
speakers to Mandarin and English (Bradley, 2012a)
indicated that Yoruba speakers also show specific advantages
over English speakers in melody discrimination. Yoruba
speakers did not perform identically to Mandarin speakers,
suggesting that all tone languages do not affect melody
perception in the same way; however, Yoruba results did not
entirely conform to expectations based on the Yoruba tone
inventory. Additional investigations are underway to
determine how differences in melody perception between
speakers of different tone languages relate to the tone
inventories of these languages, and whether task effects and
learning paradigms affect performance in a manner
consistent with the OPERA hypothesis.

Rather than a complete theory, the mapping proposed
here is a starting point for investigating the relationships
between these pitch properties, and properties may need to
be added, removed, or their relationship revised upon the
examination of additional data. It is important to note that
this is a mapping between phonetic-level properties, with the
assumption that similar phonetic properties are built from
similar acoustic properties. This is complicated by
differences in the acoustic and phonological properties of
linguistic and musical systems. For example, the shape of
pitch change, in addition to the rate of change, over the
course of a syllable may be an important cue to tone in some
languages (Chandrasekaran et al., 2007a), whereas musical
notes are typically described and notated as discrete pitch
changes, despite their complex realization by different
instruments. Indeed, Merrill et al. (2012) identified separate
cortical regions encoding gliding and discrete pitch changes.
As formulated above, the mapping between slope and
interval may not adequately represent these subtleties, and is
likely to be extended or refined through additional study
and the revision of assumptions about the languages and
music in question.

Because this mapping is mainly phonetic in nature, it is
related to but separate from the phonological/categorical
knowledge of the linguistic or musical system. That is,
musicians cannot directly transfer their knowledge of
musical intervals to tone perception, nor can tone language
speakers directly transfer their knowledge of tones to melody
perception. This is consistent with the principles of the
Shared Sound Category Learning Mechanism Hypothesis
(Patel, 2008b), which supposes that the rules and
representations of the two systems are separate, though they
are learned through the same mechanisms. Bidelman,
Gandour, and Krishnan (2011a) highlight this point by
demonstrating that, although musicians and Mandarin
speakers each show enhanced brainstem encoding of F0,
only musicians demonstrate enhanced ability to detect
mistuned melodic intervals. Although other studies have
demonstrated superior melodic perception by Mandarin
speakers, the tasks involved discrimination of one interval
type from another (e.g., minor third from major third),
something about which even untrained listeners possess
implicit knowledge, rather than the kind of precise tuning
detection to which only musicians are accustomed. Thus,
although Mandarin speakers benefit from enhanced encoding of $F_0$, their representations of musical intervals are still not as robust as the representations of musicians, who are further along the musical equivalent of the “phonetic-phonological-lexical continuity” described by Wong and Perraschione (2007). The tuning task requires greater meaning to be associated with these musical categories, which apparently only comes with music experience. Bidelman, Gandour, and Krishna (2011b) further demonstrated the importance of meaningful musical and linguistic experience by comparing the perception of musical intervals and Mandarin tones by Mandarin speakers and musicians. They found that although the FFR of both groups tracked changes in $F_0$ accurately in both lexical tone and melodic interval stimuli, the strength of representation for pitch varied with each. This likely indicates portions of the signal that are of different salience to each group in its domain of experience.

**Summary**

Pitch is an important perceptual component of both language and music. Although the surface acoustic features of lexical tone and melody are very different, they share important abstract structural features. Experience in each domain produces many parallel effects on the perceptual system. A theoretical framework for relating these parallels has been identified based on general theories of perceptual learning, and a mapping between features of lexical tone and musical melodies has been proposed that links melodic contour, interval, and tonic/key to linguistic direction, slope, and height, respectively. This framework appears to be partly consistent with many previously observed cross-domain effects, and has the potential to serve as a starting point for probing these effects further.

In order to fully elaborate this framework, future work in this area must also consider a greater range of music experience, including the diversity of musical practice in Western culture and around the world. A greater understanding of the cognitive relationship between musical and linguistic pitch would be valuable not only within the field of music cognition, but would contribute to a more general understanding of human perceptual and cognitive systems, and the place of music and language within them.

**References**


hove, m., sutherland, m. & krumhansl, c. (2010). ethnicity effects in relative pitch. psychonomic bulletin & review, 17, 310-316. doi: 10.3758/pbr.17.3.310


kras, n. & chandrasekaran, b. (2010). music training for the development of auditory skills. nature reviews neuroscience, 11(8), 599-605. doi: 10.1038/nrn2882


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), 1092-105. doi: 10.1162/jocn.2009.21077


liu, s. & samuel, a. g. (2004). perception of mandarin lexical tones when f0 information is neutralized. language and speech, 47(2), 109-138. doi: 10.1177/0023830904047020101


magne, c., schon, d. & besson, m. (2006). musician children detect pitch violations in both music and language better than nonmusician children: behavioral and electrophysiological
Pitch perception


Pitch perception


