Applied Psycholinguistics **36** (2015), 1459–1491
doi:10.1017/S0142716414000496

# Universality and language-specific experience in the perception of lexical tone and pitch

DENIS BURNHAM, BENJAWAN KASISOPA, and AMANDA REID
*University of Western Sydney*

SUDAPORN LUKSANEEYANAWIN
*Chulalongkorn University*

FRANCISCO LACERDA
*Stockholm University*

VIRGINIA ATTINA
*University of Western Sydney*

NAN XU RATTANASONE
*Macquarie University*

IRIS-CORINNA SCHWARZ
*Stockholm University*

DIANE WEBSTER
*University of Western Sydney*

Received: February 15, 2013 Accepted for publication: May 11, 2014

ADDRESS FOR CORRESPONDENCE
Denis Burnham, MARCS Institute, University of Western Sydney, Bankstown Campus Locked Bag 1797, Penrith New South Wales 2751, Australia. E-mail: denis.burnham@uws.edu.au

ABSTRACT
Two experiments focus on Thai tone perception by native speakers of tone languages (Thai, Cantonese, and Mandarin), a pitch–accent (Swedish), and a nontonal (English) language. In Experiment 1, there was better auditory-only and auditory–visual discrimination by tone and pitch–accent language speakers than by nontone language speakers. Conversely and counterintuitively, there was better visual-only discrimination by nontone language speakers than tone and pitch–accent language speakers. Nevertheless, visual augmentation of auditory tone perception in noise was evident for all five language groups. In Experiment 2, involving discrimination in three fundamental frequency equivalent auditory contexts, tone and pitch–accent language participants showed equivalent discrimination for normal Thai speech, filtered speech, and violin sounds. In contrast, nontone language listeners had significantly better discrimination for violin sounds than filtered speech and in turn speech. Together the

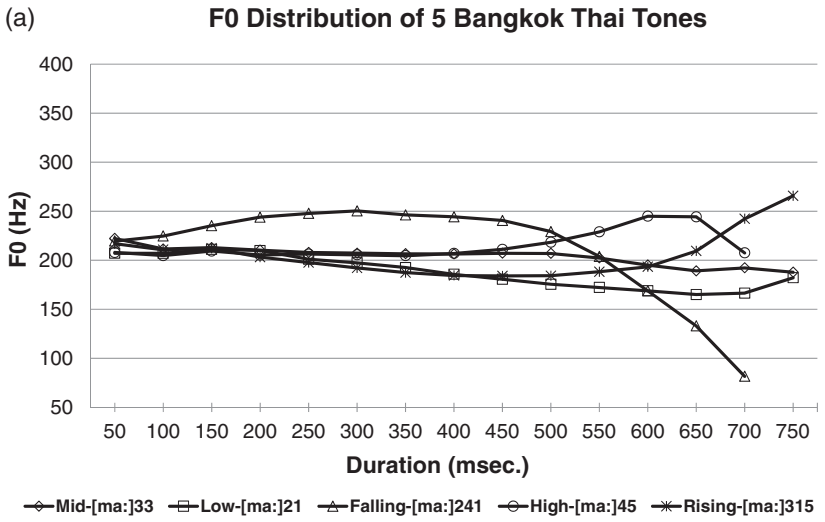
results show that tone perception is determined by both auditory and visual information, by acoustic and linguistic contexts, and by universal and experiential factors.

In nontone languages such as English, fundamental frequency (F0; perceived as pitch) conveys information about prosody, stress, focus, and grammatical and emotional content, but in tone languages F0 parameters also distinguish clearly different meanings at the lexical level. In this paper, we investigate Thai tone perception in tone (Thai, Cantonese, and Mandarin), pitch–accent (Swedish), and nontone (English) language participants. While cues other than F0 (e.g., amplitude envelope, voice quality, and syllable duration) may also contribute to some lesser extent to tone production and perception, F0 height and contour are the main distinguishing features of lexical tone. Accordingly, tones may be classified with respect to the relative degree of F0 movement over time as static (level) or dynamic (contour). In Central Thai, for example, there are five tones: two dynamic tones, [kʰǎ:]-rising tone, meaning "leg"; and [kʰâ:]-falling tone, "to kill"; and three static tones, [kʰá:]-high tone, "to trade"; [kʰa:]-mid tone, "to be stuck"; and [kʰà:]-low tone, "galangal, a root spice."

Tone languages vary in the number and nature of their lexical tones; Cantonese has three static and three dynamic tones, and Mandarin has one static and three dynamic tones. Another important variation is between tone and pitch–accent languages; in tone languages, pitch variations occur on individual syllables, whereas in pitch–accent languages, it is the relative pitch between successive syllables that is important. In Swedish, for example, there are two pitch accents that are applied to disyllabic words. Pitch Accent 1 is the default "single falling" or acute tone; for example, *anden* (single tone) [ˈa͡nd̂ɛn] meaning "duck." Pitch Accent 2 is the "double" or grave tone, which is used in most native Swedish nouns that have polysyllabic singular forms with the principal stress on the first syllable; for example, *anden* (double tone) [ˈa͡nd̂ɛn] meaning "spirit." However, while pitch accent is used throughout Swedish spoken language, there are only about 500 pairs of words that are distinguished by pitch accent (Clark & Yallop, 1990).

Figure 1 shows the F0 patterns over time of the languages of concern here (Thai, Mandarin, and Cantonese tones) and the two Swedish pitch accents. To describe the tones in these languages, both in Figure 1 and throughout the text, we apply the Chao (1930, 1947) system in which F0 height at the start and end (and sometimes in the middle) of words is referred to by the numbers 1 to 5 (1 = low frequency, 5 = high frequency), in order to capture approximate F0 height and contour.

Tone languages are prevalent; they are found in West Africa (e.g., Yoruba and Sesotho), North America and Central America (e.g., Tewa and Mixtec), and Asia (e.g., Cantonese, Mandarin, Thai, Vietnamese, Taiwanese, and Burmese). Pitch–accent languages are found in Asia (Japanese and some Korean dialects) and Europe (Swedish, Norwegian, and Latvian). Tone and pitch–accent languages comprise approximately 70% of the world's languages (Yip, 2002) and are spoken by more than 50% of the world's population (Fromkin, 1978). Psycholinguistic investigations of tone perception fail to match this prevalence. Here, we contribute



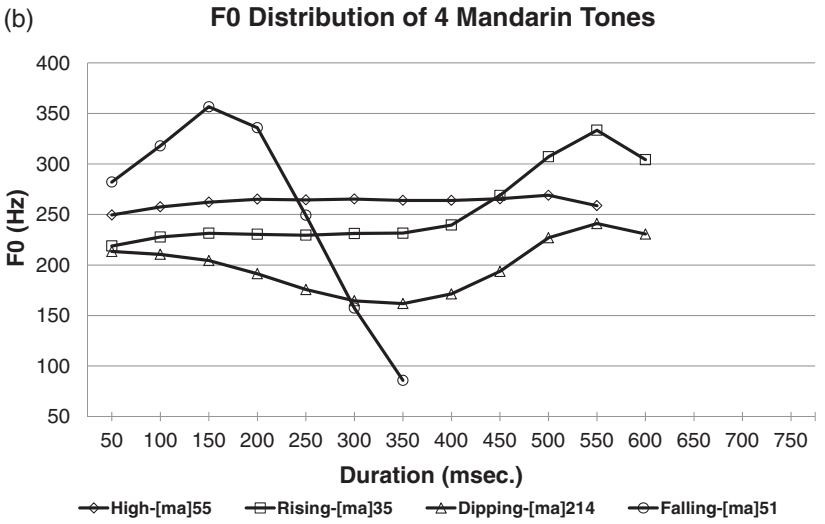
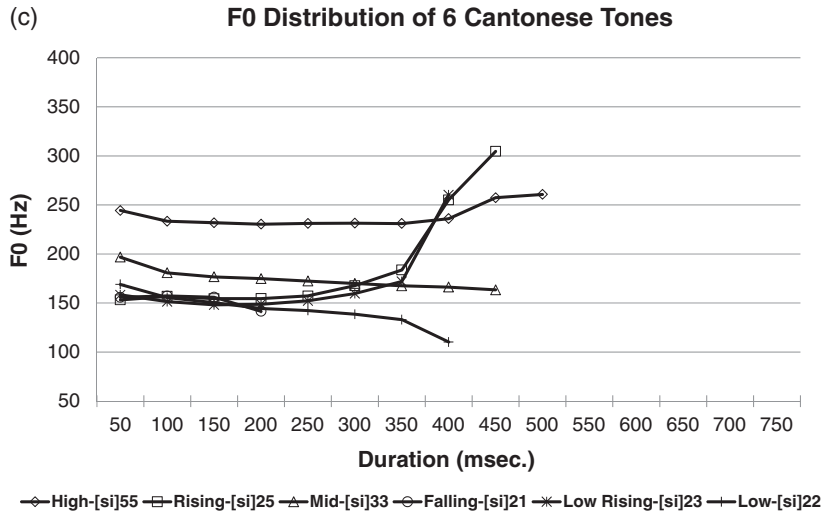




Figure 1. (a) Fundamental frequency (F0) distribution of Thai tones, based on five Thai female productions of "ma" (described by Chao values as follows: Mid-33, Low-21, Falling-241, High-45, and Rising-315). (b) F0 of Mandarin tones, based on four Mandarin female productions of "ma" (described by Chao values as follows: High-55, Rising-35, Dipping-214, and Falling-51). (c) F0 distribution of Cantonese tones, based on two Cantonese female productions of "si" (described by Chao values as follows: High-55, Rising-25, Mid-33, Falling-21, Low-Rising-23, and Low-22). (d) F0 distribution of Swedish pitch accents (across two syllables) based on three Swedish female productions for two-syllable words. Pitch Accent 1 shows the single falling F0 pattern and Pitch Accent 2 shows the double peak in F0.



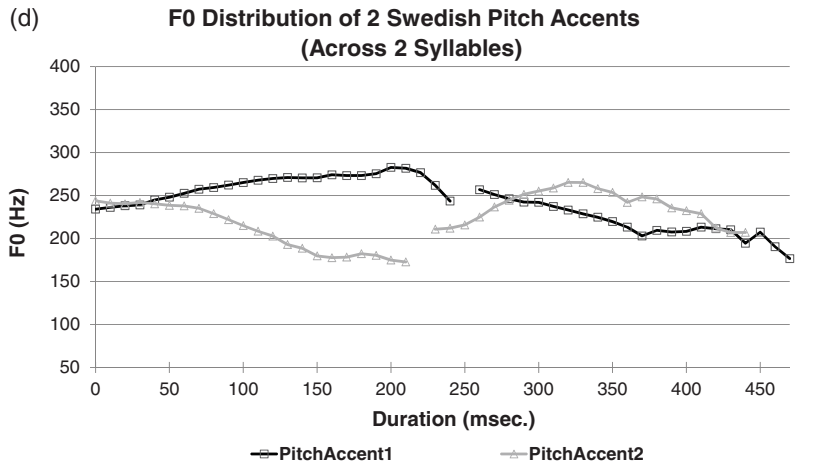




Figure 1 (*cont.*)

to redressing the balance by investigating the nature of tone perception in two experiments.

Experiment 1, a study of *cross-language* and *auditory–visual* (AV) perception, involves tests of tone discrimination in auditory-only (AO), AV, and visual-only (VO) conditions. Visual speech information is used in speech perception when available (Vatikiotis-Bateson, Kuratate, Munhall, & Yehia, 2000), and it affects perception even in undegraded listening conditions (McGurk & MacDonald, 1976). Although visual speech has been studied extensively over the

last two decades in the context of consonants, vowels and prosody (Campbell, Dodd, & Burnham, 1998), this is not the case for tone; visual speech is a necessary component of a comprehensive account of tone perception. Experiment 2 drills down to the processes of tone perception: Thai tone discrimination is tested, again within and across-languages, in three auditory contexts: speech, filtered speech, and violin sounds. By such means, we are able to draw conclusions about the relative contribution of universal and language-specific influences in tone and pitch perception. Ahead of the experiments, literature concerning perceptual reorganization for tone and the factors in auditory and AV tone perception is reviewed.

## TONE LANGUAGE EXPERIENCE AND PERCEPTUAL REORGANIZATION IN INFANCY

As a product of linguistic experience, infants' perception of consonants and vowels becomes attuned to the surrounding language environment, resulting in differential perceptual reorganization for native and nonnative speech sounds (Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995; Tsushima et al., 1994; Werker & Tees, 1984a). In addition, Mattock, Burnham, and colleagues provide strong evidence of such perceptual reorganization for lexical tone (Mattock & Burnham, 2006; Mattock, Molnar, Polka, & Burnham, 2008). Recently, it has been suggested that this occurs as young as 4 months of age (Yeung, Chen, & Werker, 2013), slightly earlier than that for vowels (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994). Mattock et al. (2008) found that 4- and 6-month-old English and French infants discriminate nonnative Thai lexical tone contrasts ([bâ] vs. [bă]), while older 9-month-olds failed to do so. Moreover, while English language infants' discrimination performance for F0 in linguistic contexts deteriorates between 6 and 9 months, there is no parallel decline in discrimination performance for nonspeech (F0-equivalent synthetic violin) contrasts (Mattock & Burnham, 2006). In contrast, Chinese infants' discrimination was statistically equivalent at 6 and 9 months for both lexical tone and violin contrasts, showing that perceptual reorganization for tone is both language specific and specific to speech. These results suggest that the absence of phonologically relevant lexical tones in English infants' language environment is sufficient to draw their attention away from lexical tone contrasts but not from nonlinguistic pitch. Experiment 2 here extends this work to adults, comparing discrimination performance of tone, pitch–accent, and nontone language adults across different tone contexts, including F0-equivalent violin contrasts.

## AUDITORY PERCEPTION OF NONNATIVE TONES AND LINGUISTIC EXPERIENCE

Studies have shown that linguistic experience (or lack thereof) with a particular native language tone set plays a role in adult listeners' auditory identification and discrimination of nonnative linguistic tones (Burnham & Francis, 1997; Lee, Vakoch, & Wurm, 1996; Qin & Mok, 2011; So & Best, 2010; Wayland & Guion, 2004). Francis, Ciocca, Ma, and Fenn (2008) posit that such perception is

determined by the relative weight given to specific tone features, which in turn is determined by the demands of the native language. Gandour (1983) showed that English and Cantonese speakers rely more on average F0 height than do Mandarin and Thai speakers, while Cantonese and Mandarin speakers rely more on F0 change/direction than do English speakers (see also Li & Shuai, 2011). Tone language background listeners usually perform better than nontone language listeners, although in some cases specific tone language experience actually results in poorer performance; for example, So and Best (2010) found that Cantonese listeners incorrectly identified Mandarin tone 51 as 55, and 35 as 214, significantly more often than did Japanese or English listeners.

There also appears to be specific effects of the number of static tones in a language. While Mandarin (one static tone) speakers generally perform better than English and French speakers on discrimination of Cantonese (three static tones) tones, English and French speakers distinguish static tones better than do Mandarin speakers (Qin & Mok, 2011). Chiao, Kabak, and Braun (2011) found that the ability to perceive the four static tones of the African Niger–Congo language, Toura, was inversely related to the number of static tones in the listeners' native first language (L1) tone system. Taiwanese (two static tones) listeners had more difficulty perceiving all static tone comparisons than did Vietnamese (one static tone) listeners or German nontone listeners. Taiwanese listeners particularly had trouble discriminating the three higher Toura tones, presumably because Taiwanese have more categories in the higher frequency region, causing more confusion. In Experiment 1 we investigated Thai tone discrimination performance of both Cantonese and Mandarin listeners, in order to determine the effect of the different number of static tones of the nonnative tone language when discriminating Thai tones. Because Cantonese has three static tones and Mandarin one and English none, it could be hypothesized that Cantonese listeners may find discrimination of Thai static tones more difficult than Mandarin listeners, and that English listeners may perform better than both of those groups on certain static tone contrasts.

In addition to the effects of language experience, it appears that there may also be a physiological bias in the registration of F0 direction. Krishnan, Gandour, and Bidelman (2010) showed that across tone language speakers, the frequency-following response to Thai tones is biased toward rising (cf. falling) pitch representation at the brain stem. Moreover, tone language speakers showed better pitch representation (i.e., pitch tracking accuracy and pitch strength) than did nontone (English) language perceivers; and tonal and nontonal language speakers could be statistically differentiated by the degree of their brain stem response to rising (but not falling) pitches. The authors suggest that this is due to a tone-language experience dependent enhancement of an existing universal physiological bias toward rising (cf. falling) pitch representation at the brain stem. Here we examine whether this possible bias toward rising pitch is evident in behavioral discrimination, and further, whether a similar bias is evident in the visual perception of tone. In Experiment 1 we investigate further the language-dependent and universal features of tone perception in cross-language tone discrimination and extend the investigation to visual features of tone by including AO, AV, and VO conditions.

## VISUAL FACILITATION OF TONE PERCEPTION: NATIVE LANGUAGE SPEECH PERCEPTION

Visual speech (lip, face, head, and neck motion) information is used in speech perception when it is available (Vatikiotis-Bateson et al., 2000). In a classic study, Sumby and Pollack (1954) demonstrated a 40%–80% augmentation of AO speech perception when speech in a noisy environment is accompanied by the speaker's face. Even in undegraded viewing conditions, an auditory stimulus, /ba/, dubbed onto an incongruent visual stimulus, /ga/, results in an emergent percept, /da/ (McGurk & MacDonald, 1976). Evidence of visual cues for lexical tone was first presented by Burnham, Ciocca, and Stokes (2001). Native Cantonese listeners asked to identify spoken words as one of six Cantonese words, differing only in tone in AV, AO, and VO modes, showed equivalent performance in AO and AV conditions. However, in the VO condition, tones were identified significantly better than chance under certain conditions: for tones in running speech (but not for words in isolation), for tones on monophthongal (but not diphthongal) vowels, and for dynamic (but not static) tones.

Mandarin listeners also show AV augmentation of identification of Mandarin tones in noise but not when F0 information is filtered out based on linear predictive coding (Mixdorff, Hu, & Burnham 2005), and similar results were also found for Thai (Mixdorff, Charnvivit, & Burnham, 2005). Finally, Chen and Massaro (2008) observed that Mandarin tone information was apparent in neck and head movements, and subsequent training drawing attention to these features successfully improved Mandarin perceivers' VO identification of tone.

## VISUAL FACILITATION OF TONE PERCEPTION: ACROSS LANGUAGES

AV speech perception in general may operate differently in tone and pitch–accent languages than in nontone languages. Sekiyama (1994, 1997) found that English language adults' McGurk effect perception is more influenced by visual speech than is that of their native Japanese-speaking counterparts, and that the increase in visual influence for English language perceivers emerges between 6 and 8 years (Sekiyama & Burnham, 2008; see also Erdener & Burnham, 2013). Moreover, Sekiyama also found even less McGurk effect visual influence for Chinese listeners (Sekiyama, 1997), although Chen and Hazan (2009) reported that Chinese and English perceivers use visual information to the same extent, but that English perceivers use visual information more when nonnative stimuli are presented. These studies compared tone and nontone language speakers on their use of visual information with McGurk-type stimuli; very few studies have compared such groups on visual information for tone.

Visual information appears to enhance nonnative speech perception in general (e.g., Hardison, 1999; Navarra & Soto-Faraco, 2005), and this is also the case with respect to tone. Smith and Burnham (2012) asked native Mandarin and native Australian English speakers to discriminate minimal pairs of Mandarin tones in five conditions: AO, AV, degraded (cochlear-implant-simulation) AO, degraded AV, and VO (silent video). Availability of visual speech information improved discrimination in the degraded audio conditions, particularly on tone

pairs with strong durational differences. In the VO condition, both Mandarin and English speakers discriminated tones above chance, but tone-naive English language listeners outperformed native listeners. This shows that visual speech information for tone is available to all perceivers, both native and nonnative alike, but is possibly underused by normal-hearing tone language perceivers. It is important to examine the parameters of English speakers' counterintuitive visual perception of tone by comparing English speakers' performance not only to native tone language perceivers but also to nonnative tone language perceivers.

Negative transfer from an L1 to a second language (L2) has also been reported in AV speech perception (Wang, Behne, & Jiang, 2008), so this could also possibly occur for tone language speakers' perception of nonnative tones. Visual cue use by nonnative perceivers may be affected by many factors, including the relationship between the inventories of visual cues in L1 and L2, the visual salience of particular L2 contrasts, the weighting given to visual versus auditory cues in a particular L1, possible visual bias triggered by the expectation that the speaker is nonnative, adverse conditions such as degraded audio, and even individual speaker and perceiver visual bias (Hazan, Kim, & Chen, 2010). In Experiment 1 here, such possibilities are explored in a new context: AO, VO, and AV AX discrimination of minimal pairs of syllables differing only on lexical tone.

## EXPERIMENT 1: AV PERCEPTION OF LEXICAL TONE

For Experiment 1, our research questions and hypotheses were as follows:

1. How does language background affect the auditory discrimination accuracy of Thai tones? It is hypothesized that there will be graded auditory performance with a rank order of Thai > (Mandarin, Cantonese, and Swedish) > English. This is based on the relative experience with tone, and Thai tones specifically, afforded by the participant's language background. However, on contrasts involving static tones, it is possible that an English > Mandarin > Cantonese pattern may be obtained (see Chiao et al., 2011; Qin & Mok, 2011). It is also hypothesized that contrasts involving rising tones will be better discriminated than other contrast pairs for all language groups (see Krishnan et al., 2010).
2. Can Thai tones be discriminated more accurately than chance on the basis of visual information alone, and how might this interact with language background? Is there any indication of a bias toward rising tones in VO conditions and are there any particular tone contrasts for which there seems to be more visual information? It is hypothesized that English speakers will outperform the native Thai speakers (see Smith & Burnham, 2012). Whether they also outperform nonnative tone (Mandarin and Cantonese) and/or pitch–accent (Swedish) speakers will have implications for the nature of any nonnative visual tone perception advantage.
3. Is there visual augmentation for Thai tones in noisy conditions, and how does this interact with language background? It is hypothesized that there will be visual augmentation for the perception of Thai tones in noisy conditions (given visual information for tone, Burnham et al., 2001) and how this manifests across language groups will have implications for how readily perceivers access visual information in adverse circumstances.

## Method

### Participants.

THAI. Thirty-six native speaking Thai listeners (21 females) were recruited from the University of Technology, Sydney (UTS) and various language centers in Sydney, Australia. The average age was 29 years ($SD = 4.0$), and the average duration of time in Australia prior to testing was 2 years ($SD = 2.8$).

MANDARIN. Thirty-six native-speaking Mandarin listeners (25 females) were recruited from UTS, the University of Western Sydney (UWS), and the University of Sydney. Most came from the People's Republic of China with 2 participants from Taiwan. The average age was 25 years ($SD = 3.7$), and the average duration of time in Australia prior to testing was 1 year ($SD = 0.7$).

CANTONESE. Thirty-six native-speaking Cantonese listeners (23 females) were recruited from UWS, UTS, the University of New South Wales, other language centers in Sydney, Australia, and the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. The average age was 22 years ($SD = 1.9$). All participants recruited in Australia came from Hong Kong ($N = 29$), and the average duration of time in Australia prior to testing was 1.5 years ($SD = 2.2$).

SWEDISH. Thirty-six native-speaking Swedish listeners (20 females) were recruited from Stockholm University, Sweden. The average age was 27 years ($SD = 9.8$).

AUSTRALIAN ENGLISH. Thirty-six native-speaking Australian English listeners (28 females) were recruited from UWS. The average age was 24 years ($SD = 7.3$).

None of the participants had received any formal musical training longer than 5 consecutive years, with the exception of 11 members of the Swedish group, because it proved difficult to recruit Swedish participants while including this criterion. All participants were given a hearing test, and all had normal hearing (at or under 25 dB at each of 250, 500, 1000, 2000, 4000, and 8000 dB). All non-Thai participants were naive to the Thai language. All participants gave informed consent to participate in the experiment, and received AUD$30 or equivalent financial compensation for participation, or received course credit.

*Design.* For each of the five language groups, a 2 (interstimulus interval [ISI]) × 2 (initial consonant × noise) × 3 (vowels × mode of presentation) × (10 [tone pairs] × 4 [AB conditions]) × 2 (repetitions) design was employed in an AX task. The first between-subjects factor was an ISI of 500 or 1500 ms. The second between-subjects factor was a nested combination of *initial consonant sound* (/k/ vs. /$k^h$/) and the presence or absence of auditory background *noise* (clear vs. noise). The third between-subjects factor was a nested combination of the *vowel sounds* (/a/, /i/, and /u/) and the *mode of presentations* (AV, AO, and VO) of the stimuli. In each language group, half of the participants were assigned the 500 ms and the

other half to the 1500 ms ISI condition. Again within each ISI condition, half of the participants were assigned to tests with initial /k/ in the clear condition and /kʰ/ in noise and the other half with initial /kʰ/ in clear and /k/ in noise. Within each resultant subgroup, one-third of the participants were assigned to AV stimuli with vowel /a/, AO stimuli with vowel /i/, and VO stimuli with vowel /u/; the second group to AV stimuli with /i/, AO stimuli with /u/, and VO stimuli with /a/; and the last group to AV with /u/, AO with /a/, and VO with /i/. The net result was that there was systematic variation across consonants and vowels in order to provide external validity of tone discrimination results across voiced versus voiceless consonants and the /a/, /i/, and /u/ vowels.

However, the most important between-subjects manipulations were auditory noise versus clear, and Mode (AO, VO, and AV), and the consonant and vowel factors will not be reported or discussed further. Similarly, preliminary analyses indicated that there was no significant main effect of ISI, nor any ISI × Language two-way interactions, and ISI will therefore not be reported or discussed further in Experiment 1.

The three within-subjects factors were the type of *Tone Pair*, *Sequence of Presentation*, and *Repetition of Condition*. In the stimulus language, Thai, there are 3 static tones, and 2 dynamic tones (see Figure 1a). Of the 10 possible tone pairings, there are 3 StaticStatic tone pairs, 1 DynamicDynamic tone pair, and 6 StaticDynamic tone pairs. The three StaticStatic pairs are Mid-Low (ML), Mid-High (MH), and Low-High (LH). The DynamicDynamic pair is Rising-Falling (RF). The 6 StaticDynamic pairs are Mid-Falling (MF), Mid-Rising (MR), Low-Falling (LF), Low-Rising (LR), High-Falling (HF), and High-Rising (HR). Therefore, for the first within-subjects factor, Tone Pair, there were 10 levels. Regarding the next within-subjects factor, Sequence of Presentation, each of the 10 possible tone pairs was presented four times to control order and same/different pairings; that is, given a pair of tone words, A and B, there were two different trials (AB and BA), and two same trials (AA and BB) trials. For the final within-subjects factor, Repetition, all of these stimuli were presented twice. The exemplars of particular phones (see below) were varied randomly even within same (AA and BB) trials to ensure that the task involved discrimination between tone categories rather than discrimination of exemplars within those tone categories.

The $d'$ scores were calculated for each of the 10 tone pairs in each condition, given by $d' = Z(\text{hit rate}) - Z(\text{false positive rate})$, with appropriate adjustments made for probabilities of 0 (=.05) and 1 (=.95). A *hit* is defined as a "different" response on an AB or BA trial and a *false positive* as a response on an AA or BB trial.

*Stimulus materials.* Stimuli consisted of 6 Thai syllables (/ka:/, /ki:/, /ku:/, /kha:/, /khi:/, and /khu:/) each carrying each of the 5 Thai tones. The resultant syllables are either words ($n = 21$) or nonwords (=9).[1] The 30 syllables were recorded in citation form by a 27-year-old native Thai female. The speaker was required to read aloud in citation form syllables displayed on a screen. The productions were audio-visually recorded in a sound-treated booth using a Lavalier AKG C417 PP microphone and a HDV Sony HVR-V1P video camera remotely controlled with Adobe Premiere software. The digital audiovisual recordings were stored at 25

video frames/s and 720 × 576 pixels, and 48-kHz 16-bit audio. Many repetitions were produced by the speaker, but only three good quality exemplars of each of the 30 syllables were selected for the experiment. Recordings were labeled using Praat, and the corresponding videos were automatically cut from Praat TextGrids using a Matlab® script and Mencoder software and stored as separate video files. To ensure that the whole lip gesture of each syllable was shown in its entirety, 200 ms of the original recording was retained at the boundaries when each syllable video file was cut. Sound level was normalized, and all videos were compressed using the msmpeg4v2 codec.

There were two auditory noise conditions: noisy and clear. In noise conditions, a multitalker Thai speech babble track was played simultaneously with the presentation of each stimulus, with a signal to noise ratio of –8 dB. Note that the VO mode also contained background babble noise in the noise condition.

*Procedure.* Participants were tested individually in a sound-attenuated room or a room with minimal noise interference on individual Notebook Lenovo T500 computers running DMDX experimental software (see Forster & Forster, 2003). They were seated directly in front of a monitor at a distance of 50 cm, and auditory stimuli were presented via high-performance background noise canceling headphones (Sennheiser HD 25-1 II), connected through an EDIROL/Cakewalk UA-25EX USB audio interface unit. Auditory stimuli were presented at a comfortable hearing level (60 dB on average). The visual component of the stimuli (i.e., the face of the Thai speaker) was presented at the center of the computer screen in an 18 cm wide × 14.5 cm high frame. For the AO condition, a still image of the talker was shown.

Each participant received a total of 480 test trials, 2 (noise/clear) × 3 (AO/VO/AV) × 10 Tone Pairs × 4 AB Conditions × 2 Repetitions split into 2 test files (for blocked testing of clear and noise stimuli). Each noise or clear test file was split into 2 120-trial test blocks. In each block, 40 trials in each mode (AO, VO, and AV), made up of 10 tone pairs and 4 AB orders, were presented randomly, and across blocks different repetitions were used. Block order was counterbalanced between subjects. At the start of each test file, 4 training trials were presented: 1 AV, 1 AO, and 1 VO trial in a training session, then another AV trial placed at the start of the test session as the decoy or warm-up trial.

Participants were instructed to listen to and watch a sequence of two videos of a speaker pronouncing syllables and to determine whether the two tones were the same or different by pressing, as quickly and accurately as possible, the right shift key if they perceived them to be the same and the left shift key if different. The time-out limit for each test trial was 5 s. If a participant failed to respond on a particular trial, he or she was given one additional chance to respond in an immediate repetition of the trial. Participants were given breaks in between each block.

## Results

*Overall analyses.* Mean $d'$ scores for each language group are shown separately for AO/AV and VO scores by noise condition (averaged over individual tone

contrasts) in Figure 2. The auditory (AO and AV scores) and VO data were analyzed separately. The alpha level was set to 0.05, and effect sizes are given for significant differences. To examine auditory speech perception (AO and AV) and visual augmentation (AO vs. AV), a 5 (language: Thai, Mandarin, Cantonese, Swedish, and English) × 2 (noise [noisy/clear]) × 2 (mode [AO/AV]) analysis of variance (ANOVA) was conducted on AO and AV scores. To examine visual speech perception, a 5 (language: Thai, Mandarin, Cantonese, Swedish, and English) × 2 (noise [noisy/clear]) ANOVA was conducted on VO scores. In each analysis, four orthogonal planned contrasts were tested on the language factor: English versus all others (i.e., nontonal English vs. the tone and pitch–accent languages); Thai + Cantonese + Mandarin versus Swedish (i.e., tone languages vs. pitch–accent language); Thai versus Cantonese + Mandarin (i.e., native vs. nonnative tone languages); and Cantonese versus Mandarin. All two- and three-way interactions were also tested. In addition, in order to test whether VO speech perception was above chance for each language group, *t* tests were conducted comparing VO $d'$ scores (overall or split on the noise factor if warranted) against chance ($d' = 0$).

*Auditory + visual augmentation ANOVA (AO and AV scores).* The results of the 5 (language: Thai, Mandarin, Cantonese, Swedish, and English) × 2 (noise [noisy/clear]) × 2 (mode [AO/AV]) ANOVA showed significantly better performance overall in clear audio than in noisy audio, $F\ (1, 170) = 805.06$, $p < .001$, partial $\eta^2 = 0.83$, and significantly better performance overall in AV than AO conditions, $F\ (1, 170) = 17.66$, $p < .001$, partial $\eta^2 = 0.09$. A significant interaction between mode and noise, $F\ (1, 170) = 30.20$, $p < .001$, partial $\eta^2 = 0.07$, indicated that across language groups, visual augmentation was present only in noise (means, $AV_{noise} = 2.2$, $AO_{noise} = 1.9$), not in clear audio ($AV_{clear} = 3.6$, $AO_{clear} = 3.7$; see Figure 2a, b).

Turning to the language factor, English language participants performed significantly worse ($M_{English} = 2.6$) overall than all other groups combined, $F\ (1, 170) = 12.95$, $p < .001$, partial $\eta^2 = 0.07$ ($M_{Thai} = 3.2$, $M_{Mandarin} = 2.9$, $M_{Cantonese} = 2.7$, $M_{Swedish} = 2.9$). There was no significant difference between the combined tone languages and the Swedish pitch–accent groups, that is, no tone versus pitch–accent language effect. However, there was significantly better performance by the native tone (Thai) than the nonnative tone (Mandarin and Cantonese) language speakers, $F\ (1, 170) = 12.58$, $p = .001$, partial $\eta^2 = 0.07$, with no overall difference between the nonnative tone language groups, Cantonese and Mandarin. There were no significant Mode × Language, Noise × Language, or Mode × Noise × Language interactions (see Figure 2a, b), showing that, despite some indication that Thai (but not other) participants were better in clear AV than AO, augmentation of AO tone perception by the addition of visual information was consistent across all five language groups.

*Visual speech ANOVA (VO scores).* The results of the 5 (language: Thai, Mandarin, Cantonese, Swedish, and English) × 2 (noise [noisy/clear]) ANOVA showed a significant effect of noise, $F\ (1, 170) = 9.01$, $p = .003$, partial $\eta^2 = 0.05$, with

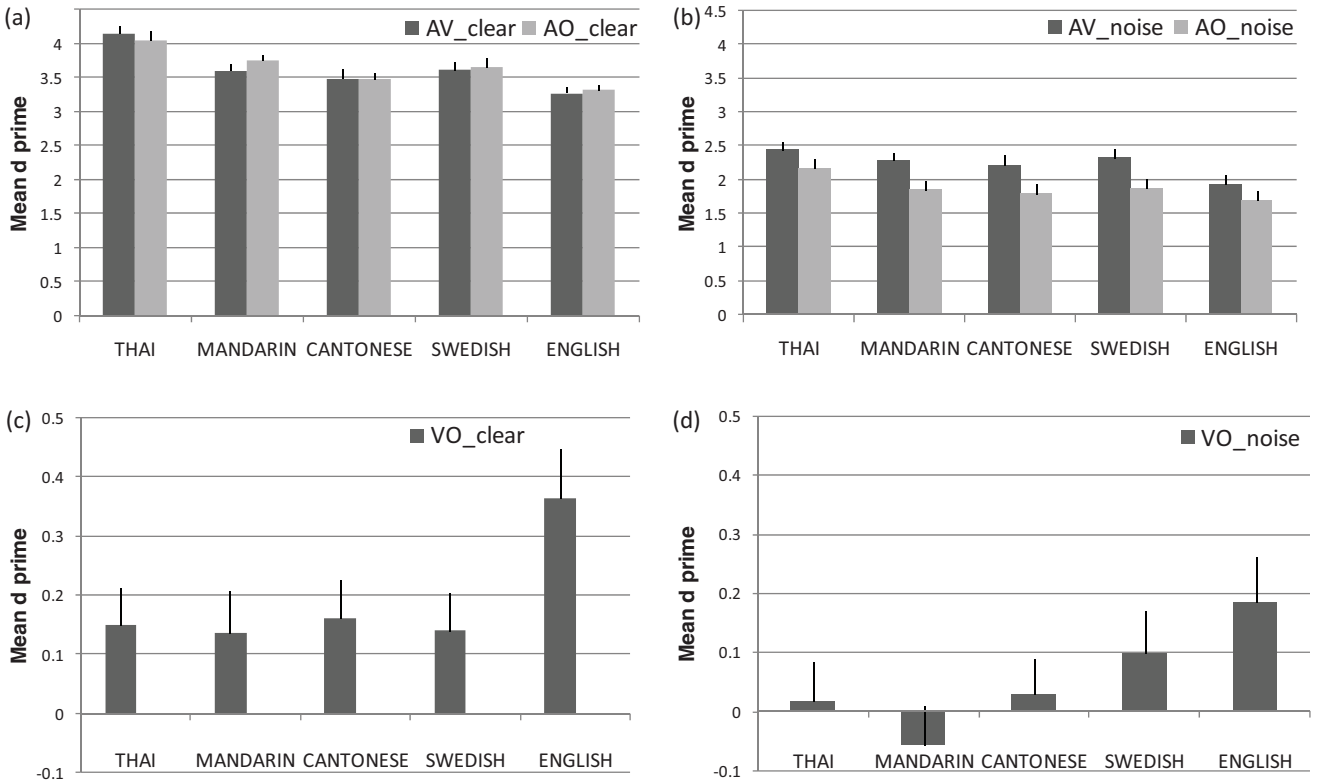



Figure 2. Mean $d'$ scores (bars are standard errors) for each language group, shown separately for auditory–visual/auditory-only (AV/AO) (a) clear and (b) noise and visual-only (VO) (c) clear and (d) noise, averaged over individual tone contrasts. Note the different scales used for AV/AO and VO figures, and standard errors are comparable across conditions.

Table 1. *Discriminability of each tone pair for auditory only (AO) scores in clear, visual only (VO) in clear, and auditory–visual (AV-AO) augmentation in noise*

| Type | Tone Pair | Mean $d'$ Scores | | |
|---|---|---|---|---|
| | | AO Clear | VO Clear | Augmentation in Noise |
| DynamicDynamic | RisingFalling | 4.0 | 0.6 | 0.8 |
| StaticDynamic | HighRising | 3.9 | 0.1 | 0.4 |
| StaticDynamic | MidRising | 3.9 | 0.4 | 0.3 |
| StaticDynamic | LowRising | 3.8 | 0.3 | −0.1 |
| StaticDynamic | HighFalling | 3.9 | −0.1 | 0.2 |
| StaticDynamic | MidFalling | 2.4 | 0.1 | 0.3 |
| StaticDynamic | LowFalling | 3.5 | 0.0 | 0.5 |
| StaticStatic | MidLow | 3.5 | 0.1 | 0.2 |
| StaticStatic | MidHigh | 3.9 | 0.3 | 0.7 |
| StaticStatic | LowHigh | 3.7 | 0.1 | 0.2 |

*Note:* AO clear $SE = 0.1$, VO clear $SE = 0.1$, augmentation in noise $SE = 0.1$–$0.2$.

better performance overall in clear than in noisy audio ($M_{\text{clear}} = 0.19$, $M_{\text{noisy}} = 0.05$). Because the noise was auditory, not visual, and because this was the VO condition, this is likely to be due to factors such as distraction or channel capacity. With respect to language effects, the English group performed significantly better than all of the other groups combined ($M_{\text{English}} = 0.27$, $M_{\text{tone/pitch–accent}} = 0.08$), $F\ (1, 170) = 11.75$, $p = .001$, partial $\eta^2 = 0.06$ (see Figure 2c, d). There were no significant differences on any of the other language contrasts, or any Noise × Language interactions.

*Visual speech t tests against chance (VO scores).* For VO scores, *t* tests against chance showed that VO performance was significantly better than chance in clear audio, but not in noisy audio, for Thai, $t\ (35) = 2.39$, $p = .022$, Cantonese, $t\ (35) = 2.39$, $p = .022$, and Swedish, $t\ (35) = 2.14$, $p = .039$, participants. For English participants, VO performance was significantly better than chance in *both* clear, $t\ (35) = 4.32$, $p < .001$, and noisy audio, $t\ (35) = 2.39$, $p = .022$. For Mandarin participants, VO performance was not significantly better than chance in either noise condition.

The following are key points:

- There was significantly better tone perception overall in AV compared with AO conditions, and this augmentation was consistent across all five language groups.
- In VO conditions, the English group performed significantly better than all other groups combined and better than chance in both clear and noisy audio. Mandarin listeners did not perceive VO tone better than chance in any condition.

*Relative discriminability of each tone pair.* Table 1 shows the discriminability of each tone pair in the three mode/conditions that resulted in the highest scores: AO scores in clear audio, VO in clear audio, and AV-minus-AO (A-V) augmentation in

noise (it was only in noise that an augmentation effect was obtained). Three single factor (tone pair) repeated measures ANOVAs were conducted, one for each of the three Mode × Condition interactions, with scores collapsed across languages. Nine planned orthogonal contrasts were tested:

1. Pairs only involving dynamic versus those also involving static tones (Dynamic-Dynamic vs. StaticStatic + StaticDynamic);
2. StaticStatic versus StaticDynamic;
3. within StaticStatic: MH + LH versus ML;
4. MH versus LH;
5. within StaticDynamic: pairs involving rising versus pairs involving falling tones (HR + MR + LR vs. HF + MF + LF);
6. HR + MR versus LR;
7. HR versus MR;
8. HF + MF versus LF; and
9. HF versus MF.

Only those contrasts on which a significant difference was found are reported.

AO IN CLEAR AUDIO. In clear AO, RF (i.e., the one and only DynamicDynamic pair) was significantly more discriminable than all other pairs combined, $F$ (1, 179) = 18.58, $p < .001$, partial $\eta^2 = 0.09$. StaticStatic pairs were slightly but significantly more easily discriminated than StaticDynamic pairs, $F$ (1, 179) = 5.70, $p = .018$, partial $\eta^2 = 0.03$, mainly due to a marked difficulty with the MF pair. In addition, StaticDynamic pairs involving the rising tone ($M = 3.9$) were significantly more discriminable than those involving the falling tone ($M = 3.5$), $F$ (1, 179) = 69.52, $p < .001$, partial $\eta^2 = 0.28$. Due to difficulty with the MF pair, LF was significantly more easily discriminated than HF + MF combined, $F$ (1, 179) = 6.05, $p = .015$, partial $\eta^2 = 0.03$, and HF significantly more easily discriminated than MF, $F$ (1, 179) = 116.15, $p < .001$, partial $\eta^2 = 0.39$. Among the StaticStatic pairs, ML was significantly more difficult to discriminate than MH + LH combined, $F$ (1, 179) = 6.86, $p = .01$, partial $\eta^2 = 0.04$. Overall these results may be described as {**DynamicDynamic** > (**StaticStatic** [MH = LH] > ML)} > {**StaticDynamic-rise** [LR = MR = HR]} > {**StaticDynamic-fall** [LF > (HF > MF)]}.

VO IN CLEAR AUDIO. In VO, RF (i.e., the DynamicDynamic pair) was significantly more discriminable than all other pairs combined, $F$ (1, 179) = 16.45, $p < .001$, partial $\eta^2 = 0.08$. StaticDynamic pairs involving the rising tone were significantly more discriminable than those involving the falling tone, $F$ (1, 179) = 7.44, $p = .007$, partial $\eta^2 = 0.04$. The MR pair was discriminated significantly more easily than the HR, $F$ (1, 179) = 5.41, $p = .021$, partial $\eta^2 = 0.03$. Thus there was a {**DynamicDynamic**} > {(**StaticStatic** = **StaticDynamic** [**StaticDynamic-rise** (LR = (MR > HR))] > [**StaticDynamic-fall** (LF = HF = MF)]} pattern of results.

AUGMENTATION (AV-AO) IN NOISY AUDIO. For visual augmentation (AV-AO), RF (the DynamicDynamic pair) showed significantly more augmentation than all

other pairs combined $F$ (1, 179) = 11.65, $p$ = .001, partial $\eta^2$ = 0.06. Among the StaticStatic pairs, MH had significantly more visual augmentation than LH, $F$ (1, 179) = 7.77, $p$ = .006, partial $\eta^2$ = 0.04, while among the StaticDynamic pairs, LR had significantly less visual augmentation than HR + MR combined, $F$ (1, 179) = 5.32, $p$ = .022, partial $\eta^2$ = 0.03. Thus, there was an overall pattern of {**DynamicDynamic**} > {**StaticStatic** [MH > LH] = ML} = {**StaticDynamic** [**StaticDynamic-rise** (LR < (MR > HR)] = [**StaticDynamic-fall** (LF = HF = MF)]}.

The following are key points:

- In clear AO and VO conditions, RF was significantly more discriminable than all other pairs. Further, other pairs involving the rising tone were significantly more discriminable than those involving the falling tone.
- RF was also associated with significantly more visual augmentation in noise than all other pairs.

*Language differences in discriminability of tone contrasts.* Thirty individual single-factor between-group ANOVAs were conducted on the language factor for all 10 tone contrasts in the three sets of greatest interest: AO in clear, VO in clear, and augmentation (AV-minus-AO) in noise. As above, four orthogonal planned contrasts were tested on the language factor: English versus all others; Thai + Cantonese + Mandarin versus Swedish; Thai versus Cantonese + Mandarin; and Cantonese versus Mandarin. Figure 3 sets out the results of these analyses, with $F$ values only shown for contrasts that were significant at 0.05 or beyond ($F_c = 3.90$). It also incorporates graphical representations of the mean $d'$ scores on each individual contrast for each language group, in AO clear, VO clear, and augmentation (AV-AO) in noise.

AO IN CLEAR AUDIO. The common direction of language differences for AO in clear audio was Thai > Mandarin + Cantonese (native tone better than nonnative tone language groups), Mandarin > Cantonese, and (Thai + Mandarin + Cantonese + Swedish) > English (English worse than all other groups combined). On the MF pair, the Thai group was markedly better than other groups, all of whom had particular difficulty with this pair. We note that this pattern was the same for AV clear, but not the same for AO or AV in noise (see Figure 4a, c, d), for which LR was the most difficult contrast for all groups.

VO IN CLEAR AUDIO. For VO in clear audio scores, language differences are predominantly evident on pairs involving the midtone, with the English group showing an advantage over other groups, particularly on MF and MR. The Cantonese group found MH particularly easy to discriminate. (However, note that these patterns were not the same in VO noise, presumably due to some distraction; see Figure 4b.)

VISUAL AUGMENTATION (AV-AO). For visual augmentation in noise, language differences are predominantly evident on pairs involving the rising tone, although the pattern here was not consistent. On LR, the nontone language groups (Swedish

| | ML | MF | MH | MR | LF | LH | FH | LR | FR | HR |
|---|---|---|---|---|---|---|---|---|---|---|
| **AO CLEAR** | | | | | | | | | | |
| *E vs all others* | | | | 7.66 E<all | | | 8.69 E<all | 9.31 E<all | 10.61 E<all | |
| *T+M+C vs S* | | | | | | | | | | |
| *T vs M+C* | | **38.55 T>MC** | | 4.38 T<MC | | 6.39 T>MC | 6.09 T>MC | | | 4.74 T>MC |
| *M vs C* | 4.37 M>C | | | | | | 4.19 M>C | 9.81 M>C | | |
| **VO CLEAR** | | | | | | | | | | |
| *E vs all others* | | 7.91 E>all | | 4.27 E>all | | | | | | |
| *T+M+C vs S* | | | | | | | | | | |
| *T vs M+C* | | 4.47 T>MC | | | | | | | | |
| *M vs C* | | | 6.60 M<C | | | | | | | |
| **AV-AO AUGMENTATION in NOISE** | | | | | | | | | | |
| *E vs all others* | | | | | | | | | 6.29 E<all | 7.55 E<all |
| *T+M+C vs S* | | | | | | | | **11.80 TMC<S** | 10.03 S<TMC | 4.87 S<TMC |
| *T vs M+C* | | | | | | | | 3.94 T<MC | | |
| *M vs C* | | | | | | | | | | |

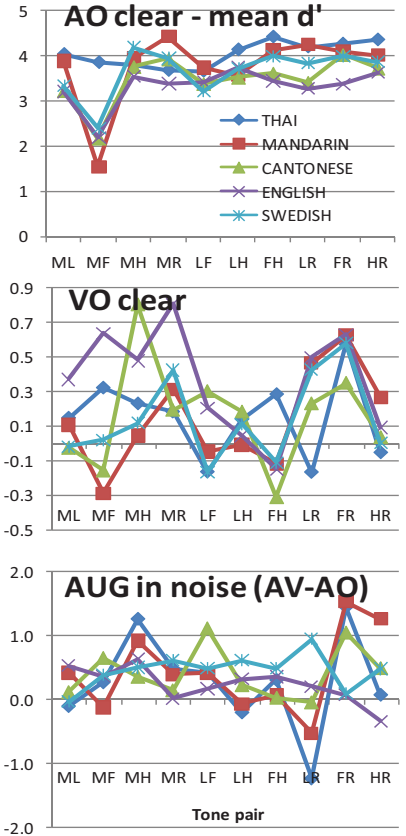



Figure 3. (Color online) The $F$ (1, 175) values for single factor language ANOVAs conducted for all 10 tone contrasts in auditory only (AO) in clear, visual only (VO) in clear, and auditory–visual/AO (AV-AO) augmentation in noise. Blank cells indicate $p > .05$, no shading indicates $p < .05$, light shading indicates $p < .01$, and dark shading indicates $p < .001$. T, Thai; M, Mandarin; C, Cantonese; E, English; S, Swedish.

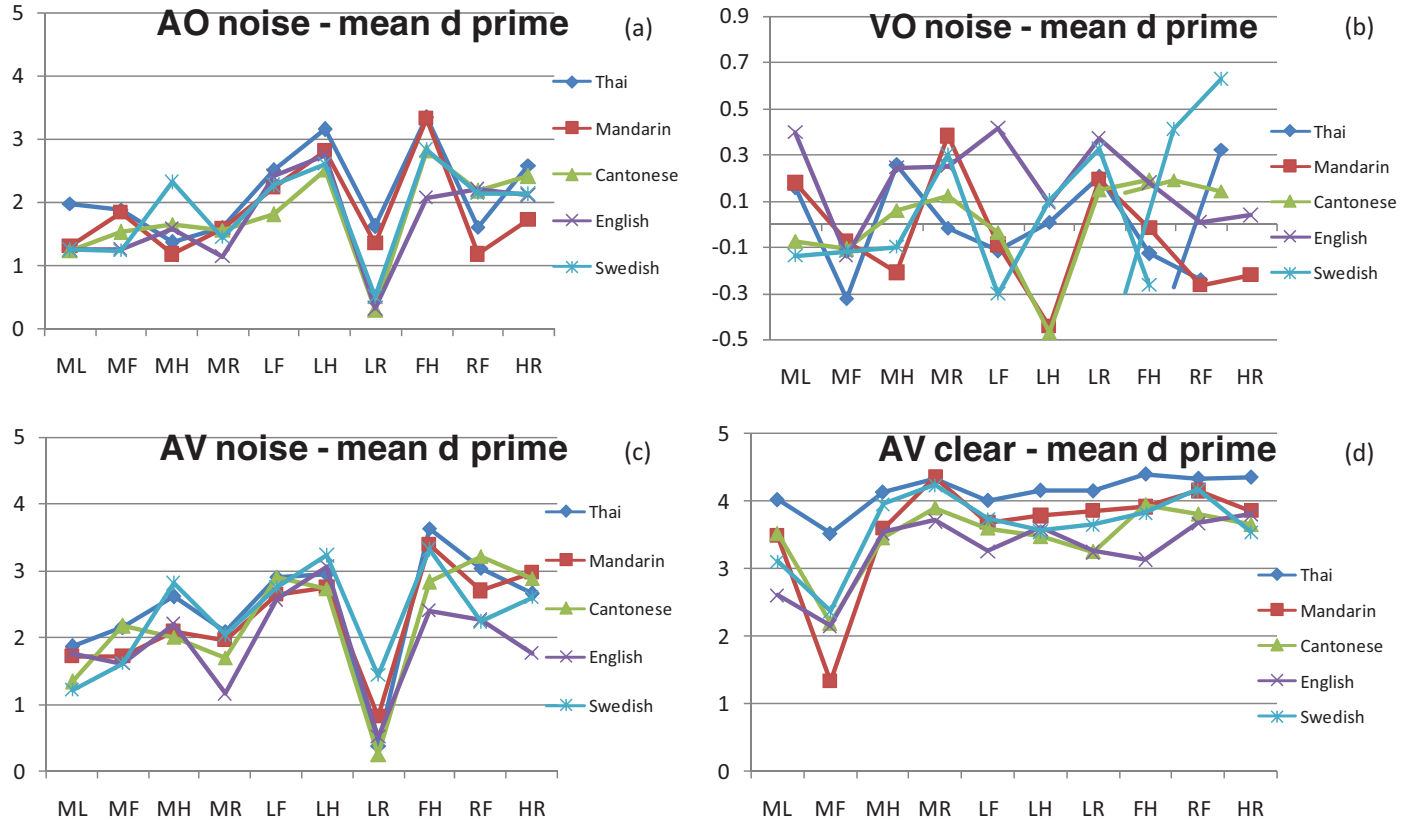



Figure 4. (Color online) Mean $d'$ scores (by tone pair) for each language group, shown separately for auditory-only (AO) noise, auditory–visual (AV) noise, visual-only (VO) noise, and AV clear. Note the different scales used for AV/AO and VO figures.

and English) showed more augmentation (and significantly so for the Swedish) than the tone groups. Within these three groups, augmentation was greater for the nonnative (Mandarin and Cantonese) than the native Thai tone groups (visual information appeared to disrupt discrimination for the tone groups for a contrast that was already particularly difficult in noise). In contrast, for FR and HR, the tone language groups showed relatively high augmentation, and the English and Swedish did not. The pattern of low performance on LR, high performance on FR was most extreme in the Thai group, and was evident across both VO and augmentation performance measures.

The following are key points:

- In clear AO conditions, the usual direction of language differences was Thai better than all others, Mandarin better than Cantonese (particularly on pairs involving a static tone) and English worse than all other groups combined (particularly on pairs involving the rising tone). There was particular difficulty with MF for all nonnative groups.
- For clear VO conditions, language differences are predominant on pairs involving the midtone, with the English group showing an advantage over other groups.
- For AV minus AO, language differences are predominant on pairs involving the rising tone, although the pattern here was not consistent.

### *Discussion*

This experiment provides clear and strong evidence for (a) language-general visual augmentation (AV > AO) of tone perception regardless of language background; (b) language-specific facilitation of tone perception by tone or pitch–accent language experience in AO and AV conditions, and nontone experience in VO conditions; and (c) effects on performance due to particular tone contrasts.

*Visual augmentation.* Over and above any other effects, in the auditory noise condition there was augmentation of AO perception of tone by the addition of visual information (AV > AO). This was unaffected by language experience whatsoever; visual augmentation was equally evident across all five language groups. These results provide strong evidence that there is visual information for lexical tone in the face, and that this can be perceived and used equally by native tone language listeners, nonnative tone language listeners, pitch–accent listeners, and even nonnative, nontone language listeners. Thus, the augmentation of tone perception by visual information is independent of language-specific experience.

*Language experience.* With respect to *auditory (AO and AV) conditions,* (a) experience with the lexical use of pitch in tone languages (Mandarin and Cantonese) or a pitch–accent language (Swedish) facilitates tone perception in an unfamiliar tone language, and (b) there is a separate advantage for perceiving tone in one's own language. Thus our hypothesis was supported; the experiential effects of tone language experience in AO and AV in clear and noisy audio conditions can be characterized as {**Tone** [native > no-native] = **Pitch Accent**} > {**Nontone**}.

For *VO perception of tone*, there are also language experience effects, but in more or less the opposite direction. Thai, Cantonese, and Swedish listeners all perceived VO tone better than chance in clear, but not noisy, audio (the latter presumably due to some cross-modal distraction). Mandarin listeners did not perceive VO tone better than chance in any condition. This appears to be in conflict with the findings by Smith and Burnham (2012), in which Mandarin participants did perceive VO tone better than chance in the VO clear condition. However, in that study, Mandarin participants were tested on their native Mandarin tones, rather than nonnative Thai tones that would likely create a more difficult task here. In addition, compared with other tone languages, Mandarin has greater durational differences between tones, and this may well lead to greater reliance on an acoustic strategy when perceiving unfamiliar tones. We do note, though, that the difference in VO performance between the Mandarin and Cantonese groups was not significant in the ANOVA.

English language listeners perceived VO tone better than chance in both clear and noisy audio. Comparison across language groups showed that nonnative nontone language English participants significantly outperformed all the other four groups, supporting our hypothesis. This English superiority was particularly evident on pairs involving the midtone. This could reflect perception of the midtone as the "norm" for English speakers; English speakers could be highly familiar with the visual cues associated with this tone. In contrast, there could be physically very few visual cues for the midtone with those for other tones standing out as distinctive in comparison. Future research is needed to elucidate this further.

This superior VO performance by English over tone language users confirms and extends results with Mandarin VO tone perception reported by Smith and Burnham (2012). There it was only possible to say that English language perceivers outperformed tone language (Mandarin) perceivers, so involvement of the foreign speaker effect, in which participants attend more to visual information when faced with a foreign rather than a native speaker (Chen & Hazan, 2009; Fuster-Duran, 1996; Grassegger, 1995; Kuhl, Tsuzaki, Tohkura, & Meltzoff, 1994; Sekiyama & Burnham, 2008; Sekiyama & Tohkura, 1993), could not be ruled out. Here, however, because there was no significant difference between the Thai and the (combined) nonnative groups (Mandarin and Cantonese), nor between these tone language groups and the pitch–accent Swedish group, the English > all lexical pitch language groups superiority cannot be due to a general foreign speaker effect. There appears to be something special about English language or nontone language experience that promotes visual perception of tone, as will be discussed further in the General Discussion.

*Tone contrast effects.* There were a number of effects specific to particular tones and tone—tone combinations. It is of interest that the MF contrast, which was a particularly difficult contrast auditorily for all nonnative groups, was the one on which the English superiority was greatest in the VO clear condition. However, this did not appear to assist the English listeners on the MF contrast in the AV condition. Because this superiority for English listeners was found in VO, but not in AV > AO augmentation, it is possible that English listeners are less able to integrate auditory and visual tone information as effectively as are native tone language and pitch–accent listeners (despite integration of consonant information as evidenced

by the McGurk effect among English listeners; McGurk & MacDonald, 1976). Similarly, the Cantonese group were relatively good at discriminating the MH contrast in VO, but this did not assist them in AV > AO augmentation.

In the AO condition, Cantonese participants performed significantly worse than Mandarin participants on ML, HF, and LR contrasts. All three of these pairs involve at least one static tone so, as hypothesized, this may be due to the greater number of static tones in Cantonese than Mandarin. In addition, there were no tone contrasts in AO for which the Cantonese performed significantly better than the Mandarin group. These results may reflect more confusion for the Cantonese group due to their additional categories for native static tones; existing static tones may act as perceptual magnets yielding poor performance (Chiao et al., 2011; Kuhl, 1991). In contrast to our hypothesis, there were no tone contrasts in AO on which the English were significantly better than the other groups combined, even on StaticStatic contrasts, which only involve static tones (see Qin & Mok, 2011).

In clear AO and VO conditions, tone pairs involving the rising tone were generally more easily discriminated than other pairs, supporting our hypothesis. This concurs with suggestions from frequency-following responses that rising tones are more salient than falling tones (Krishnan et al., 2010), and that there may be a physical bias regarding sensitivity toward F0 direction. Krishnan et al. (2010), also using Thai tones, suggested that tone language listeners (Thai and Mandarin) have developed more sensitive brain stem mechanisms for representing pitch (reflected by tracking accuracy and pitch strength) than nontone (English) language perceivers. Further, tonal and nontonal language listeners can be differentiated (using discriminant analysis) by their degree of response to rising (but not falling) pitches in the brain stem. That is, while there may be a universal bias toward rising tones across all language groups, the degree to which this is activated and expressed may depend on tone language experience. Our results support this; here the advantage for rising tones in AO was less evident when there was no tone language experience (as the English group performed significantly more poorly on LR, RF, and MR compared with the other groups), but further research is required to test the generality of this effect. The poorer performance by the English speakers on RF and some StaticDynamic tone contrasts is also in line with the fact that Cantonese and Mandarin speakers rely more on F0 change/direction in perception than do English speakers (Gandour, 1983).

Along with the better overall performance across language groups in VO on pairs involving the rising tone, it is noteworthy that the RF contrast was the most easily visually discriminable, and associated with most visual augmentation in noise. While this is intuitively reasonable and in accord with the Burnham et al. (2001) results for dynamic versus static tone identification in Cantonese, the exact visual cues involved are not clear; it is likely that they lie in rigid head movement and laryngeal movements (Chen & Massaro, 2008) rather than nonrigid facial movements (Burnham et al., 2006). The RF contrast was the most easily discriminated contrast in the AO condition also, so there appears to be a general effect at play.

The results of Experiment 1 provide information about the role of language experience in the perception of tone. There is evidence for universal language-*general* augmentation of tone perception by visual information and differential language-*specific* effects on the perception of tone (and particular tone contrasts)

in AO, VO, and AV conditions. In Experiment 2 we address more the *mechanisms* by which language experience affects tone perception.

## EXPERIMENT 2: THE EFFECTS OF LINGUISTIC EXPERIENCE ON TONE AND PITCH PERCEPTION

Experiment 2 investigates how the mechanisms of perceiving tone linguistically versus nonlinguistically might differ across listeners with different language backgrounds. Auditory (AO) Thai tone contrasts were modified, while keeping F0 constant, into two different nonspeech formats: low-pass filtered speech and violin sounds. Two different ISIs were used (500 and 1500 ms), which have been posited to force different levels of processing of speech stimuli (Werker & Logan, 1985; Werker & Tees, 1984b), with the 1500 ms ISI presumably involving deeper processing and more reliance on long-term memory.

Again, a same–different AX task was employed, and participant groups were similar to those in Experiment 1. There were four groups: native tone language speakers, Thai; nonnative but nevertheless tone language speakers, Cantonese; nonnative pitch–accent language speakers, Swedish; and nonnative, nontone language speakers, English. Only Cantonese nonnative tone language speakers were included because (a) in Experiment 1 Cantonese and Mandarin results were similar and (b) further analysis of Experiment 1 and other related data revealed discrimination of Thai tones predicted categorization for Cantonese but not Mandarin listeners.

For Experiment 2, our research questions and hypotheses were as follows:

1. How does processing tones linguistically versus nonlinguistically (in filtered speech and violin contexts) differ across language backgrounds? Based on Mattock and Burnham, (2006), it is hypothesized that English listeners will be better able to discriminate the same F0 patterns when they are presented in a nonspeech (violin or filtered speech) than a speech context, while there should be no difference for native Thai speakers or for the nonnative tone language and pitch–accent groups.
2. How is the pattern of relative accuracy for linguistic and nonlinguistic conditions affected by processing at different ISIs for each of the language groups?

### *Method*

*Participants and design.* A total of 192 adults (48 native Thai, 48 native Cantonese, 48 native Swedish, and 48 native English speakers) were tested in a Language Background (Thai, Cantonese, Swedish, and English) × ISI (500 and 1500 ms) × Tone Type (speech, filtered speech, and violin) design with repeated measures on the last factor. Half the participants in each language group were tested at each ISI, and within these subgroups, approximately half the participants were males and half females. For each group, the mean age and range were as follows: Thai: 20.9, 17–30 years; Cantonese: 20.6, 17–34 years; English: 22.0, 17–40 years. Although no Swedish age data were recorded, these were all university undergraduates, as in the other three age groups. None of the Swedish or

English speakers had ever received instruction in a tone language (other bilingual experience was not an exclusion criterion). Expert musicians were excluded from the study. (For more on musicians' tone perception with these stimuli, see Burnham, Brooker, & Reid, 2014.)

*Stimuli.* Three stimulus sets were created, speech, filtered speech, and violin, each comprising three duration-equated exemplars of each of the five Thai tones. The original speech stimuli were recorded from a female native Thai speaker using the syllable [pa:] to carry the five tones: rising [pă:], high [pá:], mid [pa:], low [pà:], falling [pâ:]. These 15 (5 tones × 3 exemplars) speech sounds were then used as a basis for the filtered speech and the violin stimuli.

The filtered speech stimuli were created by digitally low-pass filtering the speech sounds to remove all frequencies above 270 Hz. This reduced the upper formant information while leaving the F0 intact.

The violin stimuli were used because the violin can both maintain a continuous sound and reproduce rapid pitch changes (e.g., the pitch dynamics of the Thai falling tone, which covers approximately 1.5 octaves in a short space of time). A professional violinist listened extensively to the speech recordings and then reproduced approximately 25 exemplars of each tone on the violin. From these, the final 3 music exemplars for each tone were selected based on careful comparison (using the Kay Elemetrics CSL analysis package) of the pitch plots of the original lexical tone and the violin sounds, with due regard to and control of, duration. Across the 5 Tones × 3 Exemplars, the frequency range for speech was 138–227 Hz. In contrast, the frequency range for violin stimuli was higher at 293–456 Hz (in musical terms, between about D4 and A4; middle C is C4, 261 Hz and the lowest note on a violin is G3, 196 Hz). Although frequencies played were not conventional musical notes, the sound was recognizable as a violin. Figure 5 shows the F0 tracks of corresponding speech, filtered speech, and violin stimuli.

*Apparatus.* The experiment was conducted in parallel at the University of NSW (English and Cantonese speakers) and Chulalongkorn University (Thai speakers) on identical portable systems. An in-house program, MAKEDIS, was used to control presentation and timing of the sounds and record responses and reaction times. An attached response panel contained a "same" and a "different" key, and a set of colored feedback lights that were used during the training phase. At Stockholm University, Swedish participants were tested on an equivalent system.

*Procedure.* Each participant completed three AX discrimination tasks, identical except for the stimulus type: speech, filtered speech, or violin. In each, the participant first listened to a 1-min "context" tape (a woman conversing in Thai, a concatenation of filtered speech excerpts, and a violin recording of Bach's Crab Canon, respectively). Participants then completed a task competence phase, in which they were required to respond correctly on four simple auditory distinctions, two same and two different presentations of *rag* and *rug* [ɹæg, ɹʌg]. Two

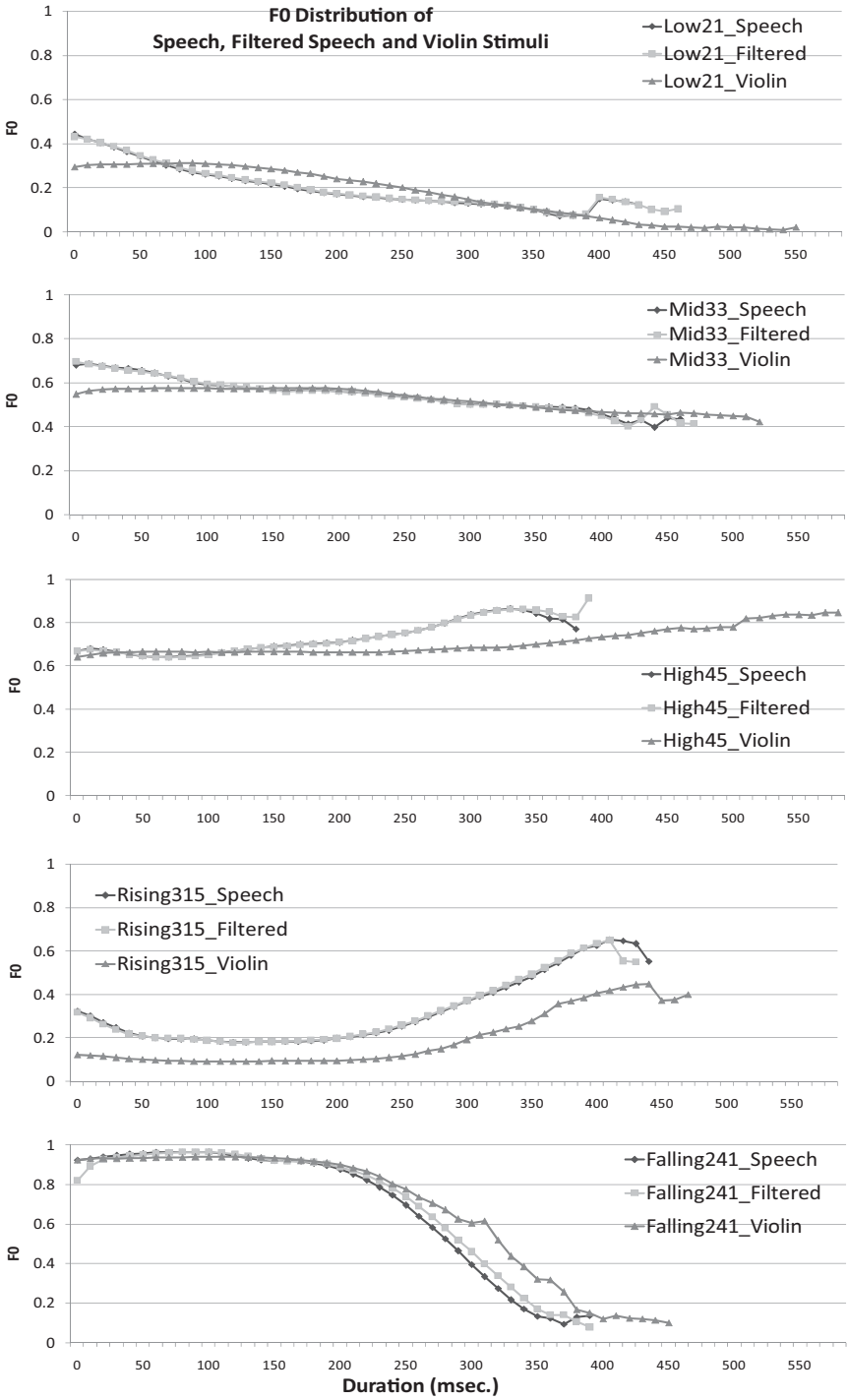



Figure 5. Fundamental frequency distribution of speech, filtered speech, and violin stimuli on each Thai tone, shown with normalized pitch.

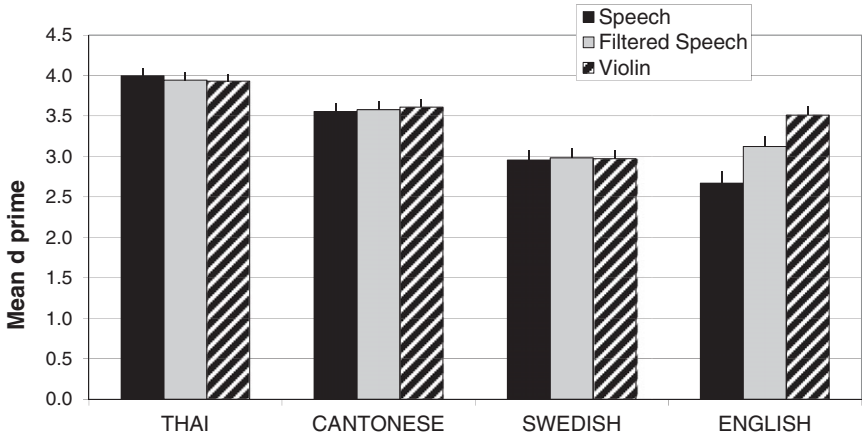



Figure 6. Thai, Cantonese, Swedish, and English speakers' mean d′ scores for speech, filtered speech, and violin tone stimuli (bars indicate standard errors).

40-trial test blocks were then given, each with 5 of the possible 10 different contrast pairs presented in the first block, and the other 5 in the second block. The order of presentation of blocks was counterbalanced between subjects. For each contrast pair, each of the four possible Stimulus × Order combinations (AA, BB, AB, and BA) were presented twice. Participants were required to listen to stimulus pairs and respond by pressing either the same or the different key within 1000 ms. (Due to some program differences, the maximum response time for Swedish participants was 1500 ms.)

Finally, participants completed two Likert rating scales, one on the similarity of each of the three sound types to speech (1 = *not at all like speech*, 7 = *exactly like speech*), and another on the similarity to music. These confirmed that for all participant groups, speech was perceived as speech, violin as music, and filtered speech as neither predominantly speech nor music.

### *Results*

Mean *d′* scores are shown in Figure 6. Scores were analyzed in a 4 (language: Thai, Cantonese, Swedish, and English) × 2 (ISI: 500 and 1500 ms) × 3 (stimulus type [speech, filtered speech, and violin]) ANOVA, with repeated measures on the last factor. Planned orthogonal contrasts tested on the language factor were: lexical pitch language (tone or pitch accent) experience versus no lexical pitch language experience (Thai + Cantonese + Swedish vs. English); tone versus pitch–accent experience (Thai + Cantonese vs. Swedish); and native versus nonnative tone experience (Thai vs. Cantonese). The planned orthogonal contrasts tested on the stimulus type factor were speech versus nonspeech (filtered speech + violin) and filtered speech versus violin. All two- and three-way interactions were also tested.

There was no significant overall effect of ISI, $F$ (1, 186) = 2.88, $p$ = .09, for $d'$, and ISI did not significantly interact with any other factor: language or stimulus type.

Tone and pitch–accent language speakers (Thai, Cantonese, and Swedish) combined ($M$ = 3.5) performed significantly better than did English speakers ($M$ = 3.1), $F$ (1, 186) = 15.61, $p$ < .001, partial $\eta^2$ = 0.08. The Thai and Cantonese groups ($M$ = 3.8) also performed significantly better than the Swedish group ($M$ = 3.0), $F$ (1, 186) = 54.21, $p$ < .001, partial $\eta^2$ = 0.23; and the Thai group ($M$ = 4.0) better than the Cantonese ($M$ = 3.6) group, $F$ (1, 186) = 8.93, $p$ = .003, partial $\eta^2$ = 0.05. Thus, the pattern of results was **[(Thai > Cantonese) > Swedish] > (English),** showing separate effects of native versus nonnative tone experience, tone versus pitch–accent language experience, and lexical pitch versus no lexical pitch language experience.

Discrimination performance on speech ($M$ = 3.3) was generally inferior to that on nonspeech performance ($M$ = 3.5), $F$ (1, 186) = 8.43, $p$ = .004, partial $\eta^2$ = 0.04, but this was qualified by a significant interaction of Thai + Cantonese + Swedish versus English both with speech/nonspeech, $F$ (1, 186) = 26.57, $p$ < .001, partial $\eta^2$ = 0.12, and with filtered/violin, $F$ (1, 186) = 9.34, $p$ = .003, partial $\eta^2$ = 0.05. Figure 6 shows that while English language listeners were worse than all other groups at perceiving tone contrasts in speech, their performance markedly improved as the stimuli become less speechlike, **Speech < Filtered < Violin**. In contrast, there were negligible differences across stimulus types for the other three (tone or pitch–accent) language groups, and post hoc tests showed no significant differences across stimulus types for the tone and pitch–accent groups: Cantonese: $F$ (2, 94) = 0.140, $p$ = .870; Thai: $F$ (2, 94) = 0.199, $p$ = .820; Swedish: $F$ (2, 98) = 0.022, $p$ = .978. In contrast, there were significant differences for the English group: $F$ (2, 94) = 21.827, $p$ < .001, partial $\eta^2$ = 0.317.

The following are key points:

- For English language listeners, performance markedly improved as the tone stimuli become less speechlike.
- There were negligible differences across speech, filtered speech, and violin for the other (tone or pitch–accent) language groups.

### *Discussion*

The results of Experiment 2 confirm and extend those of Experiment 1 regarding the role of linguistic experience in lexical tone perception, and they also shed light on the mechanisms by which this occurs. There is a three-step graded effect of linguistic experience on the perception of lexical tone. First, experience with lexical pitch (tone or pitch–accent) languages, Cantonese or Swedish, provides an advantage (over nontone language English listeners) for discriminating foreign lexical tones. Second, more specific experience with a tone language, Cantonese, versus that with a pitch–accent language, Swedish, provides an added advantage for the discrimination of foreign lexical tones. We note that this was more obvious in this experiment than in Experiment 1, in which, in auditory conditions using only tone in a speech context, Swedish participants did just as well as the full tone

language Cantonese participants. Here, possibly the nonspeech conditions act to reduce overall performance more so for pitch–accent than full tone language participants; further research with other pitch–accent language groups (such as Japanese) is required. Third, over and above these language-general effects, there is yet more, a native language advantage such that native tone language, Thai, listeners perform better than nonnative tone language, Cantonese, listeners.

Turning to the *mechanisms* of cross-language tone perception, for the nontone nonnative English language group, but none of the other groups, discrimination accuracy improves as an inverse function of the speechlike nature of the sounds: speech < filtered < violin. This supports our hypothesis that lack of lexical tone experience has a specific linguistic effect rather than a more general effect on pitch perception, and along with the Mattock and Burnham (2006) results, suggests that this specific effect originates in the perceptual reorganization for lexical tone occurring in infancy (Mattock & Burnham, 2006; Mattock et al., 2008; Yeung et al., 2013). Our results also suggest that the sequelae of early reorganization are difficult but not impossible to overcome; changing from a speech to a nonspeech context allows the nontone language listener to be freed from the constraints of processing tone at a phonemic (or phonetic) level, so that in the linguistically unencumbered violin context, they discriminate well the F0 levels and contours present in (Thai) lexical tones. It is possible then that continued practice in such nonspeech contexts (e.g., musical training) may overcome this earlier perceptual reorganization, and Burnham et al. (2014) have found evidence for this. In contrast, within each of the lexical pitch experience listener groups (Thai, Cantonese, and Swedish), there is no difference between their discrimination accuracy ($d'$) for pitch contrasts in speech, filtered speech, and violin contexts; in other words, no attenuation or facilitation of perception of F0 as a function of listening in a linguistic context.

However, there are two riders to our results. First, there was no effect of ISI, so English language-listeners' performance was improved only by release from a linguistic stimulus context, not by reducing memory constraints (from 1500 to 500 ms ISI) and the consequent disengagement of phonemic storage categories (Werker & Logan, 1985; Werker & Tees, 1984b). Second, performance in the violin context cannot be considered the upper psychoacoustic level against which attenuation due to linguistic experience can be gauged, for there are significant reductions in the violin $d'$ levels from Thai to Cantonese and English to Swedish. It is possible that this is due to the within-subjects, speech, filtered, violin manipulation, engendered by the speech context carrying over to the violin; this awaits further investigation. However, there is evidence that tone language experience also augments pitch perception (e.g., Bidelman, Hutka, & Moreno, 2013; Wong et al., 2012).

## GENERAL DISCUSSION

Together, the results of these two experiments show that experience with a tone and even a pitch–accent language facilitates the discrimination of tones in an unfamiliar tone language, compared to discrimination performance by nontone language speakers. This experiential influence is graded across three levels: facilitation due to tone or pitch–accent language experience over nonlexical-pitch language

experience; facilitation due to tone language over pitch–accent language experience (Experiment 2); and facilitation due to experience with the target than another tone language. In addition to the specific advantage imbued by lexical pitch experience on perception of an unfamiliar tone system, there are at least two clearly universal aspects of tone perception and another two that may prove to be universal.

First, listeners of all language backgrounds are able to use visual information for tone because there was better tone discrimination across the board for AV in noise than AO in noise. There is no evidence here that this information is also used in clear listening, but this is not surprising because studies of AV vowel and consonant perception also reveal the most visual influence in auditory noise conditions (Sekiyama & Burnham, 2008; see Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). What is most intriguing is whether AV influences in tone perception can be observed in clear auditory conditions in experimental paradigms akin to those giving rise to the McGurk effect (McGurk & McDonald, 1976).

Second, over and above and regardless of the AV advantage in noise, nontone language (English) perceivers were better at using VO information than were all the other language groups. This may be viewed from two angles: it may well be that learning a tone or pitch–accent language entails learning about the predominance of acoustic cues for tone (and pitch) with relatively less perceptual weight placed on visual cues for tone in clear auditory conditions (Sekiyama & Burnham, 2008); or English language perceivers' superior visual perception of tone could be because English is a nontone language, and perceivers attend to whatever cues are available to pick up unfamiliar lexical tone differences, whereas tone and pitch–accent perceivers rely more on the acoustic cues that they know to be powerful correlates of tone identity. If so, then it appears that visual perception of tone is universal and that all tone and pitch language perceivers underuse this information.

However, there does seem to be some linguistic aspect to the English VO superiority. Their VO perception of tone is especially good for contrasts involving the midtone, which is arguably the norm in English speech, and might serve as a language-specific baseline for these comparisons. Thus, the English VO superiority could partly be because English speakers are accustomed to using F0 information to perceive sentence-level intonation that deviates from the midtone norm (e.g., questions vs. statements) in English. However, there is no direct or indirect evidence to support this, and it may be more pertinent to note that English language perceivers use visual speech information more than do some other language groups (e.g., Japanese; Sekiyama, 1994, 1997). This begins between 6 and 8 years and has been suggested to occur because English phonology has a large number of auditorily confusable single phonemes and clusters coupled with a high degree of visual distinctiveness (Sekiyama & Burnham, 2008). It is interesting that the visual information in English is better used by children who are also good at focusing on native language sounds (Erdener & Burnham, 2013).

Thus, this VO superiority by English speakers could indicate universal sensitivity to visual speech information that is underused by tone and pitch–accent language speakers, and/or it could be due to the nature of the English language. Although further research is indicated, tone language speakers underuse visual information under normal circumstances, presumably due to their learned reliance

on auditory information for tone. Nevertheless, tone and pitch–accent language users appear to be better at *integrating* auditory and visual information for tone than are nontone language speakers, so research on the implication that they should be more prone to McGurk-like effects with tone is warranted. It will also be of interest to investigate (a) whether hearing-impaired tone language users better use visual tone information than do hearing tone language users; and (b) whether the use of visual information for tone can be trained in specific target groups (e.g., young tone language recipients of a cochlear implant).

Third, a potentially universal aspect of tone perception is that certain tones are perceived more easily than others. There are two aspects of this. In Experiment 1, the DynamicDynamic tone contrast (RF) was consistently perceived more easily than other contrast types in all contexts: AO, VO, and in AV > AO advantage. This suggests that tone *contour* provides the best information across modalities and that articulation of dynamic tones perhaps has more obvious visual concomitants than that for static tones. This is consistent with the Burnham et al. (2001) result of better VO tone identification of dynamic than static tones by Cantonese speakers. It is also consistent with suggestions that visual information for tone may be carried in small rigid movements of the head (Burnham et al., 2006) that are due to the small movements of the cricothyroid muscle that control the larynx when pitch is varied (Yehia, Kuratate, & Vatikiotis-Bateson, 2002).

Fourth, following from these specific aspects of tone perceptions, the last possible universal aspect of tone perception is that tone pairs involving the Thai rising tone are generally more easily discriminated than other tone pairs. This suggests that over and above any effects of tone or pitch–accent experience, rising tones are more salient than falling tones and that there may be a physical bias regarding sensitivity to F0 direction (Krishnan et al., 2010).

The mechanisms of cross-language tone perception are informed by this research. Tone and pitch–accent language speakers (Thai, Cantonese, and Swedish) perform equally well across listening contexts, speech, filtered speech, and violin, suggesting that for them lexical tone perception in speech occurs unhindered and as a natural extension of psychoacoustic pitch perception. However, nontonal English language speakers' perception of pitch in speech is attenuated below its usual psychoacoustic level (as represented in filtered speech and violin). This is presumably a product of nontonal language speakers' early perceptual reorganization (Mattock & Burnham, 2006; Mattock et al., 2008) that entails reduction of attention to irrelevant cues (in this case, pitch level and contour of individual syllables) and maintenance and focus of attention on other relevant cues. These results can be taken as evidence for specialized language-specific *linguistic* representation of pitch in adults.

We have uncovered language-general and language-specific aspects of lexical tone perception. The investigations were quite specific; only one aspect of speech (pitch) was examined, and only one aspect of pitch in speech at that (pitch on individual syllables and not prosody, sentential stress etc.). Accordingly, future studies may reveal that unlike attenuation of pitch discrimination in lexical tone, nontone language speakers have no attenuation for more global pitch patterns (intonation) in sentences. Further research is required to examine the various types of linguistic and emotional prosody in both nontone and tone languages

(Hirst & Di Cristo, 1998; Thompson, Schellenberg, & Husain, 2004). In addition, the results bear on perception only. Further studies are required to ascertain whether these results might also be reflected in the *articulation* of nonnative tones. Finally, violin sounds were used as a tool here, and there are various aspects of speech–music relations with respect to pitch and tone perception that might be examined in further studies. The speech–music parallel is more explicit in lexical tones, and tone languages provide a rich vein of information for exploring the nature of speech–music cognitive and neural relations (see Burnham et al., 2014).

Together the results show that there is a range of information available for the perception of tone, and both universal factors and participants' language background determine how this information is used. There is better use of auditory information by tone and pitch–accent language speakers, better use of VO information by nontonal perceivers, and visual augmentation for speakers of all language backgrounds. Further, changing from a speech to a nonspeech context allows the nontonal perceiver to be freed from the constraints of processing tone information at a phonemic (or phonetic) level, leading to better use of the available auditory information. Tone perception is determined by both auditory and visual information, by acoustic and linguistic contexts, and by universal and experiential factors.

## ACKNOWLEDGMENTS

For Experiment 1, we appreciate Australian Research Council funding (DP0988201 to D.B.) and assistance with data collection by Yvonne Leung, Leo Chong, members of the Thai Student Society of UTS, and Prof. Catherine McBride, Department of Psychology, Chinese University of Hong Kong. Part of the data from Experiment 1 were presented by Burnham, Attina, and Kasisopa (2011) and Reid et al. (in press). For Experiment 2, we thank Dr. Caroline Jones and Dr. Elizabeth Beach for assistance in data organization. Part of the data were presented in past conference proceedings and book chapters by Burnham, Francis, Webster, Luksaneeyanawin, Attapaiboon, Lacerda, and Keller (1996); Burnham, Francis, Webster, Luksaneeyanawin, Lacerda, and Attapaiboon (1996); and Burnham and Mattock (2006).

## NOTE

1. Although words and nonwords are likely to be processed differently, this issue is only relevant for one of the five language groups in this experiment, that is, the native Thai group, and is unlikely to impact on the overall cross-language differences that are the predominant concern here.

## REFERENCES

Best, C. T., McRoberts, G. W., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development, 18*, 339–350.

Bidelman, G., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PLOS ONE, 8*, e60676.

Burnham, D., Attina, V., & Kasisopa, B. (2011). *Auditory–visual discrimination and identification of lexical tone within and across tone languages.* Paper presented at the Auditory–Visual Speech Processing (AVSP) Conference, Volterra, Italy, September 1–2.

Burnham, D., Brooker, R., & Reid, A. (2014). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychology of Music.* Advance online publication. doi:10.1177/0305735614546359

Burnham, D., Ciocca, V., & Stokes, S. (2001). Auditory–visual perception of lexical tone. In P. Dalsgaard, B. Lindberg, H. Benner, & Z.-H Tan (Eds.), *Proceedings of the 7th Conference on Speech Communication and Technology, EUROSPEECH 2001 Scandinavia* (pp. 395–398). Retrieved from http://www.isca-speech.org/archive/eurospeech_2001

Burnham, D., & Francis, E. (1997). The role of linguistic experience in the perception of Thai tones. In A. S. Abramson (Ed.), *Southeast Asian linguistic studies in honour of Vichin Panupong* (Science of Language, Vol. 8, pp. 29–47). Bangkok: Chulalongkorn University Press.

Burnham, D., Francis, E., Webster, D., Luksaneeyanawin, S., Attapaiboon, C., Lacerda, F., et al. (1996). Perception of lexical tone across languages: Evidence for a linguistic mode of processing. In T. Bunnell & W. Isardi (Eds.), *Proceedings of the 4th International Conference on Spoken Language Processing* (Vol. 1, pp. 2514–2517). Philadelphia, PA: IEEE.

Burnham, D., Francis, E., Webster, D., Luksaneeyanawin, S., Lacerda, F., & Attapaiboon, C. (1996). Facilitation or attenuation in the development of speech mode processing? Tone perception over linguistic contexts. In P. McCormack & A. Russell (Eds.), *Proceedings of the 6th Australian International Conference on Speech Science and Technology* (pp. 587–592). Canberra, Australia: Australian Speech Science and Technology Association.

Burnham, D., & Mattock, K. (2006). The perception of tones and phones. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (Language Learning and Language Teaching Series). Amsterdam: John Benjamins.

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Haszard Morris, R., et al. (2006). *The perception and production of phones and tones: The role of rigid and non-rigid face and head motion*. Paper presented at the 7th International Seminar on Speech Production, Ubatuba, Brazil, December 1–5.

Campbell, R., Dodd, B., & Burnham, D. (Eds.). (1998). *Hearing by Eye II: Advances in the psychology of speech reading and auditory–visual speech*. East Sussex: Psychology Press.

Chao, Y.-R. (1930). A system of tone-letters. *Le Maître Phonétique, 45*, 24–27.

Chao, Y.-R. (1947). *Cantonese primer*. Cambridge, MA: Harvard University Press.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *Journal of the Acoustical Society of America, 123*, 2356–2366.

Chen, Y., & Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory–visual speech perception. *Journal of the Acoustical Society of America, 126*, 858–865.

Chiao, W.-H., Kabak, B., & Braun, B. (2011).*When more is less: Non-native perception of level tone contrasts*. Retrieved February 2, 2012, from http://ling.uni-konstanz.de/pages/home/braun/articles/Chiao.Kabak.Braun-1.pdf

Clark, J., & Yallop, C. (1990). *An introduction to phonetics and Phonology*. Oxford: Basil Blackwell.

Erdener, D., & Burnham, D. (2013). The relationship between auditory–visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology, 116*, 120–138.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, and Computers, 35*, 116–124.

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics, 36*, 268–294.

Fromkin, V. (Ed.). (1978). *Tone: A linguistic survey*. New York: Academic Press.

Fuster-Duran, A. (1996). Perception of conflicting audio–visual speech: An examination across Spanish and German. In D. G. Stork & M. E. Hennecke (Eds.), *Speech reading by humans and machines* (pp. 135–143). Berlin: Springer–Verlag.

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11*, 149–175.

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 210–213). Stockholm: Stockholm University Press.

Hardison, D. M. (1999). Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect. *Language Learning, 49*, 213–283.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52*, 996–1009.

Hirst, D., & Di Cristo, A. (Eds.). (1998). *Intonation systems: A survey of twenty languages*. Cambridge: Cambridge University Press.

Krishnan, A., Gandour, J. T., & Bidelman, G. M. (2010). The effects of tone language experience on pitch processing in the brain stem. *Journal of Neurolinguistics, 23*, 81–95.

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

Kuhl, P. K., Tsuzaki, M., Tohkura, Y., & Meltzoff, A. (1994). Human processing of auditory–visual information in speech perception: Potential for multimodal human–machine interfaces. In K. Shirai & K. Kakehi (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 539–542). Tokyo: Acoustical Society of Japan.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*, 606–608.

Lee, Y.-S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research, 25*, 527–542.

Li, B., & Shuai, L. (2011). Effects of native language on perception of level and falling tones. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 1202–1205). Hong Kong: City University of Hong Kong, Department of Chinese, Translation and Linguistics.

Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*, 241–265.

Mattock, M., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition, 106*, 1367–1381.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mixdorff, H., Charnvivit, P., & Burnham, D. K. (2005). Auditory–visual perception of syllabic tones in Thai. In E. Vatikiotis-Bateson, D. Burnham, & S. Fels (Eds.), *Proceedings of the Auditory–Visual Speech Processing International Conference* (pp. 3–8). Adelaide, Canada: Causal Productions.

Mixdorff, H., Hu, Y., & Burnham, D. (2005). Visual cues in Mandarin tone perception. In I. Trancoso, L. Oliviera, & N. Mamede (Eds.), *Proceedings of the 9th European Conference on Speech Communication and Technology* (pp. 405–408). Bonn, Germany: International Speech Communication Association.

Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*, 4–12.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception & Performance, 19*, 421–435.

Qin, Z., & Mok, P. P.-K. (2011). Discrimination of Cantonese tones by Mandarin, English and French speakers. In *The psycholinguistic representation of tone, 2011* (pp. 50–53). Hong Kong: Causal Productions.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Xu Rattanasone, N., et al. (in press). Perceptual assimilation of lexical tone: The role of language experience and visual information. *Attention, Perception & Psychology*.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*, 1147–1153.

Sekiyama, K. (1994). Difference in auditory–visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E), 15*, 143–158.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*, 73–80.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory–visual speech perception. *Developmental Science, 11*, 303–317.

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444.

Smith, D., & Burnham, D. (2012). Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. *Journal of the Acoustical Society of America, 131*, 1480–1489. doi:10.1121/1.3672703

So, C. K., & Best, C. T. (2010). Cross-language perception of nonnative tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech, 53*, 273–293.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Thompson, W. F., Schellberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion, 4*, 46–64.

Tsushima, T., Takizawa, O., Sasaki, M., Siraki, S., Nishi, K., Kohno, M., et al. (1994). *Discrimination of English /r–l/ and /w–y/ by Japanese infants at 6–12 months: Language specific developmental changes in speech perception abilities*. Paper presented at the International Conference on Spoken Language Processing, Yokohama, Japan.

Vatiktiotis-Bateson, E., Kuratate, T., Munhall, K. G., & Yehia, H. C. (2000). The production and perception of a realistic talking face. In O. Fujimura, B. D. D. Joseph, & B. Palek (Eds.), *LP'98, Item order in language and speech* (pp. 439–460). Prague: Charles University, Karolinum Press.

Wang, Y., & Behne, D., & Jiang, H. (2008). Linguistic experience and audio–visual perception of non-native fricatives. *Journal of the Acoustical Society of America, 124*, 1716–1726.

Wayland, R. P., & Guion, S. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning, 54*, 681–712.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37*, 35–44.

Werker, J. F., & Tees, R. C. (1984a). Cross-language speech perception: Evidence for perceptual reorganisation during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Werker, J. F., & Tees, R. C. (1984b). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75*, 1866–1878.

Wong, P., Ciocca, V., Chan, A., Ha, L., Tan, L., & Peretz, I. (2012). Effects of culture on musical pitch perception. *PLOS ONE, 7*, e33424.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics, 30*, 555–568.

Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language, 68*, 123–139.

Yip, M. J. W. (2002). *Tone* (Chap. 1, pp. 1–14). New York: Cambridge University Press.