Chapter 9
Multi-Modal Perception of Tone

Yue Wang, Joan A. Sereno, and Allard Jongman

Abstract This chapter surveys the role of visual cues in Chinese lexical tone production and perception, addressing the extent to which visual information involves either linguistically relevant cues to signal tonal category distinctions or is attention-grabbing in general. Specifically, the survey summarizes research findings on which visual facial cues are relevant for tone production, whether these cues are adopted in native and non-native audio-visual tone perception, and whether visual hand gestures also affect tone perception. Production findings demonstrate that head, jaw, eyebrow, and lip movements are aligned with specific spatial and temporal pitch movement trajectories of different tones, suggesting linguistically meaningful associations of these visual cues to tone articulation. Perception findings consistently show that specific facial and hand gestures corresponding to pitch movements for individual tones do benefit tone intelligibility, and these benefits can be augmented by linguistic experience. Together, these findings suggest language-specific mechanisms in cross-modal tone production and perception.

9.1 Introduction

Our understanding of the extent to which visual facial cues can aid speech communication is largely based on the perception of consonants and vowels. Studies dating back to at least the 1950s have demonstrated that segmental perception benefits from visual cues, especially when auditory distinctiveness decreases (e.g., Sumby & Pollack, 1954). Specifically, research has established that visual cues provided by speakers' facial movements, particularly those resulting from vocal tract configurations, such as lip opening, rounding, and spreading, benefit segmental speech

Y. Wang (✉)
Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada
e-mail: yuew@sfu.ca

J. A. Sereno · A. Jongman
Department of Linguistics, University of Kansas, Lawrence, KS, USA

H.-M. Liu et al. (eds.), *Speech Perception, Production and Acquisition*,
Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-15-7606-5_9

perception (Kim & Davis, 2014; Perkell, Zandipour, Matthies, & Lane, 2002; Traunmüller & Öhrström, 2007). In contrast, studies on the role of visual facial cues to the perception of prosody, including lexical tone in Chinese, did not appear until the early 2000s, and the findings have been inconclusive.

Many languages, including most Chinese languages (e.g., Cantonese, Mandarin) employ tones to convey lexical meaning, similar to the linguistic function of segmental phonemes. However, unlike phonemes, lexical tones are acoustically manifested primarily as changes in fundamental frequency (F0, perceived as pitch) as well as duration and amplitude, which are triggered by glottal and sub-glottal activities independent of vocal tract configurations (Fromkin, 1978; Howie, 1976; Lehiste, 1970; Yip, 2002). As such, although facial and even manual gestural movements have been shown to facilitate tone perception (e.g., Burnham et al., 2006; Chen & Massaro, 2008; Morrett & Chang, 2015), it is unclear whether such movements are linguistically meaningful cues to signal tonal category distinctions or general "attention-grabbing" cues.

To address these issues, this chapter provides a survey of how visual cues in Chinese tone production coordinate with acoustic tonal features and integrate with auditory cues in tone perception. In particular, the survey summarizes research findings on (1) visual facial cues (head/jaw, eyebrow, and lip movements) identified as relevant for tone production, (2) the perceptual correlates of visual facial cues in native and non-native audio-visual tone perception, and (3) the role of visual hand gestures in audio-gestural tone perception. Bringing these findings together, the chapter concludes with a discussion of the extent to which cross-modal integration of sensory-motor information in tone production and perception reflects specific linguistically motivated cues to tonal distinctions or more generic attentional cues.

9.2 Identifying Facial Cues in Tone Production

There is evidence that movements of the head, jaw, neck, eyebrows, as well as lips are associated with specific tonal or general prosodic production (Attina et al., 2010; Burnham, Ciocca, & Stokes, 2001a; Chen & Massaro, 2008; Kim, Cvejic, & Davis, 2014; Swerts & Krahmer, 2010; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Some of these movements (e.g., neck, chin) are believed to be physiologically motivated, due to movements of the laryngeal muscles that control the vocal folds when pitch is varied (Burnham et al., 2015; Yehia et al., 2002). Attempts have also been made to relate certain facial movements (e.g., head, eyebrow, lip) in terms of spatial and temporal changes in distance, direction, speed, and timing to acoustic features of tonal changes in height, contour, and duration (Attina et al., 2010; Garg, Hamarneh, Jongman, Sereno, & Wang, 2019).

9.2.1 Head and Jaw

Research has demonstrated that head movements reflect acoustic correlates in terms of F0 changes. First, the magnitude of head motion appears to be aligned with the amount of F0 variation. For instance, based on computer-vision analysis, Garg et al. (2019) found that Mandarin high-level tone (Tone 1), compared to the other tones, involves minimal head movements and low movement velocity, indicating the "level" (i.e., minimal F0 variation) nature of Tone 1. Likewise, Burnham et al. (2006) showed that head movements (e.g., nodding, tilting, rotation toward the back), as computed from the principal component analysis on kinematic sensor data, were correlated with F0 changes in Cantonese tones. These results are consistent with previous studies on prosody, that head movements are larger (and occur more frequently) in prosodic constituents with a larger amount of variance in F0 (Munhall et al., 2004; Yehia et al., 2002), for example, in sentences with strong focus (Kim et al., 2014; Swerts & Krahmer, 2010), in stressed syllables (Scarborough, Keating, Mattys, Cho, & Alwan, 2009), and in interrogative intonation (Srinivasan & Massaro, 2003).

Furthermore, it has been shown that vertical head and jaw movements are compatible with tone contour direction. Garg et al. (2019) demonstrated that upward and downward head movements follow the rising, dipping, and falling tone trajectories for Mandarin mid-high-rising tone (Tone 2), low-dipping tone (Tone 3), and high-falling tone (Tone 4), respectively. Moreover, the time taken for the movements to reach the maximum displacement is also aligned with these trajectories. Similarly, kinematic data show back and forth head movements to be correlated with F0 modulation of contour tones in general (Tones 2–4 in Mandarin, Attina et al., 2010), and a lowered jaw position correlated with the production of a low tone (Tone 3) in low vowel contexts (Shaw, Chen, Proctor, Derrick, & Dakhoul, 2014).

These patterns suggest a positive correlation between head/jaw movements and changes in F0 in the production of tonal variations. It has been speculated that head and jaw lowering or raising can be triggered by a reduction or increase in the tension of the vocal folds (movements of the cricothyroid muscle and ligaments) associated with low- or high-pitched tones, respectively (Moisik, Lin, & Eslin, 2014; Smith & Burnham, 2012; Yehia et al., 2002). However, additional quantitative data are needed to further identify the articulatory and physiological relevance of head/jaw movements in characterizing individual tonal categories and how they are associated with F0 variations.

9.2.2 Eyebrows

Eyebrow movements are also found to be associated with prosodic articulation (Kim & Davis, 2014; Swerts & Krahmer, 2010; Munhall et al., 2004; Yehia et al., 2002), although little research has focused on tone. Garg et al. (2019) showed that, similar to head movements, the spatial and temporal changes in eyebrow motion also follow the

trajectories of tone height and contour in Mandarin. Specifically, the magnitude of eyebrow displacement, as well as its movement velocity, is smaller for the level tone (Tone 1) as compared to the contour tones. For the contour tones (Tones 2, 3, and 4), eyebrow movements are aligned with the direction and timing of the rising, dipping, and falling trajectories of these tones. It should be noted that these measurements of eyebrow movements have been corrected for head motion; thus, the observed eyebrow movement patterns in tone production are not a byproduct of but rather are independent of head movements.

Despite the lack of research on tone, research examining prosodic and non-speech pitch contrasts lends some support to the patterns observed in Garg et al. (2019). Data from kinematic measures reveal larger vertical eyebrow displacement and higher peak velocity of eyebrow movements for focused (Kim et al., 2014), accented (Flecha-Garcia, 2010; Swerts & Krahmer, 2010), and stressed (Scarborough et al., 2009) words in a sentence. These results indicate that eyebrow movements may be coordinated with F0 for prosodic contrasts, although in these studies the specific connection to F0 changes (in terms of height and direction) is not straightforward or invariably evident (Ishi, Haas, Wilbers, Ishiguro, & Hagita, 2007; Reid et al., 2015). Although they did not specifically focus on prosody, Huron and Shanahan (2013) did report a causal relationship between vertical eyebrow displacement and F0 height through manipulation of eyebrow movements. By instructing the speakers to raise or lower their eyebrows to different degrees during reading, the authors found higher eyebrow placement to be associated with a higher vocal pitch.

These results reveal similar patterns of eyebrow and head movements. However, unlike the case for head motion, eyebrow movements cannot be interpreted in relation to laryngeal activities, and thus pitch. Instead, eyebrow movements in distance, direction, speed, and timing may be spatially and temporally equated with acoustic features in terms of pitch height, contour and duration, since pitch has been claimed to be audio-spatial in representation (Connell, Cai, & Holler, 2013; Hannah et al., 2017).

9.2.3 Lips

Lip movements typically signal segmental rather than prosodic contrasts, since the articulation of prosody does not rely on vocal tract configurations. Nonetheless, there has been evidence that lip movements (e.g., lip opening, lowering, inter-lip distance) may be spatially and temporally aligned with prosodic changes such as stress (Dohen & Loevenbruck, 2005; Dohen, Loevenbruck, & Hill, 2006; Scarbourough et al., 2009). For Mandarin tone production, Attina et al. (2010) reported a general correlation between lip closing and F0 irrespective of tones, as well as unique patterns for individual tones (Tones 1 and 2 only). In particular, Tone 1 was characterized by lip raising (as well as jaw advancement), suggesting a potential link between these movements and the height or the lack of contour of this high-level tone; in contrast, Tone 2 production was mainly distinguished by lip protrusion, claimed to be related

to the rising contour. In addition, for the high-falling tone (Tone 4) in Mandarin, temporal and spatial events coordinate to signal downward movement (Garg et al., 2019). Specifically, relative to the other tones, Tone 4 exhibited the longest time for the velocity of lip closing to reach maximum value and was also accompanied by the longest time for the head and the eyebrows to reach maximum lowering, which suggests that the lowering movement occurred in the later part of the tone production, corresponding to the falling F0 trajectory of this tone.

Although these studies show that certain tonal information may be carried by lip configurations, as is the case for head and eyebrows, further research is needed to pinpoint the specific movements characterizing different tone categories, and to examine if/how they correspond to changes in tone height and contour.

Taken together, these results collectively suggest that specific movements of the head, eyebrows and lips are correlated with tonal articulation and are likely coordinated with the spatial and temporal dynamics of the production of different tones. However, evidence from tone perception research is needed to determine if these facial tonal cues are indeed used to facilitate perception of categorical tonal distinctions and the extent to which perception is based on the linguistic relevance of these cues.

9.3 Audio-Visual Tone Perception

We will first discuss the use of visual (mostly, facial) cues by native perceivers and then turn our attention to non-native perceivers (including learners) whose native language is tonal or non-tonal in order to identify the linguistically relevant cues to visual tone perception.

9.3.1 Native Perceivers

Pioneering research by Burnham and colleagues first established the presence of visual facial cues for tones. Burnham et al. (2001a) tested the identification of the six tones of Cantonese. Cantonese perceivers were presented with Cantonese words in three modes: Audio-Visual (AV), in which they both saw and heard a video clip of the speaker; Audio-Only (AO), in which they heard the speaker and saw a still picture of her face; and Video-Only (VO), in which perceivers saw the video clip without any sound. Overall, there was no evidence of visual augmentation: Perceivers were found to be equally accurate in the AV mode (mean accuracy: 82.6%) and the AO mode (82.2%). Moreover, performance in the VO mode (18.6%) was at chance level. While these results suggest that visual information does not augment auditory tone perception, more detailed analyses revealed that perception in the VO mode was better than chance under certain conditions. Specifically, visual information was helpful for perceivers without phonetic training, but not those with phonetic training;

for tone carried on monophthongs, but not diphthongs; for tones spoken in a carrier phrase, but not in isolation form; and for contour tones, but not level tones. Thus, under certain circumstances, visual information did play a role. While perceivers' tone identification was not very accurate, it was significantly better than chance, indicating that there is helpful visual information in tone articulation.

Mixdorff, Hu, and Burnham (2005) replicated the basic findings of Burnham et al. (2001a) for Mandarin. They presented Mandarin Chinese perceivers with the four tones of Mandarin in AV and AO modes. Accuracy was very high (near 100%) and no difference between the AV and AO modes was observed. This absence of visual augmentation was also reported for Thai with an AX discrimination task in which the five tones were presented pair-wise in AV, AO, and VO modes (Burnham et al., 2015), and performance in the VO condition was significantly better than chance.

The findings from these studies indicate that visual cues to tone are present (performance in VO mode is better than chance) but native tonal perceivers do not additionally benefit from visual information over that provided by the auditory signal (performance in AV mode is not better than in AO mode).

However, similar to Sumby and Pollack's (1954) observation for segmental distinctions, visual information may become more prominent as auditory information becomes degraded and more difficult to access. The following section examines whether a visual benefit for tonal distinctions can be observed under auditorily challenging conditions such as the presence of background noise, hearing impairment, or when presented with non-native input.

9.3.1.1 Perception in Noise

Given that auditory tone perception has been found to be less accurate in noise (e.g., Lee & Wiener, Chap. 1 of this volume), it is conceivable that perception may benefit from complementary visual information in such adverse auditory conditions. Indeed, while Mixdorff et al. (2005) did not report any difference between the identification of Mandarin tones in AV and AO modes, an advantage for the AV mode over the AO mode became apparent when the same stimuli were presented in babble noise at different signal-to-noise ratios (SNR). Specifically, as the SNR decreased, the relative gain of the AV mode over the AO mode increased from 1.3% at −3 dB to 15.3% at −12 dB. Similar results were reported for Thai (Burnham et al., 2015), with an AV advantage only found when the stimuli were presented in multi-talker Thai babble noise at an SNR of −8 dB. Overall, these results suggest that the visual benefit stems from the early integration of acoustic and visual cues rather than additional information in the video signal per se.

9.3.1.2 (Simulated) Hearing Impairment

The influence of facial information has been well documented for segmental perception in populations with hearing loss (Campbell, Dodd, & Burnham, 1998; Grant,

Walden, & Seitz, 1998; Schorr, Fox, Wassenhove, & Knudsen, 2005). Research suggests that hearing-impaired perceivers may show a greater reliance on visual information than normal-hearing perceivers (Desai, Stickney, & Zeng, 2008; Rouger, Lagleyre, Fraysse, Deneve, Deguine, & Barone, 2007). However, it remains to be seen if this holds true for the perception of tone as well.

Smith and Burnham (2012) took a first step toward addressing this question by using simulated cochlear implant audio. That is, the audio signal was processed in a way to make it similar to that perceived by users of a cochlear implant (CI). Since CIs are poor at providing clear pitch information, CI users may rely more on visual cues to tone than normal-hearing perceivers. Mandarin stimuli were presented in an AX discrimination task in five conditions: AV, AO, VO, CI-simulated AV, and CI-simulated AO. Mandarin perceivers performed significantly better than chance in the VO condition but showed no advantage for the AV over the AO condition. However, when the acoustic signal was degraded to resemble CI speech, perceivers did significantly better in the CI-simulated AV than in the CI-simulated AO condition. These data also suggest that an impoverished audio signal encourages the use of visual information.

9.3.1.3 Directed Attention

Chen and Massaro (2008) investigated whether perceivers could be trained to pay attention to specific visual cues to Mandarin tones. Mandarin perceivers' tone identification was tested before and after training. This study focused exclusively on the VO mode. During training, participants were instructed to pay attention to mouth, head/chin movements, and especially activities of the neck. They were also allowed to use a sheet that summarized the main visual correlates (cues relating to activity of the neck and chin) of each of the four tones. Before training, perceivers' accuracy was significantly above chance at 33%. After training, their accuracy was significantly better, at 48%. Thus, perceivers' awareness of and use of visual cues to tone can be improved through specific training.

9.3.1.4 Individual Tones

So far, we have observed overall gains in performance due to the presence of visual cues with and without accompanying auditory cues. However, while certain consonants (most notably those with more anterior articulations) can benefit more from visual information than others (e.g., Jongman, Wang, & Kim, 2003), similarly, not all tones benefit equally from the presence of the speaker's face. In their study of Cantonese, Burnham et al. (2001a) reported better-than-chance performance in the VO mode only for the dynamic tones, i.e., those tones whose pitch contour exhibited movement. In Mandarin, the visual gain (AV better than AO) observed by Mixdorff et al. (2005) in babble noise occurred only for the contour tones (Tones 3 and 4). Chen and Massaro (2008) also reported better-than-chance performance in the VO mode

for Tones 2 and 3 before training; after training, performance on Tone 1 was above chance as well. In their discrimination task with all possible pairings of the Mandarin tones, Smith and Burnham (2012) found that level-contour contrasts (Tone 1-Tone 3 was most discriminable) were better discriminated than contour–contour contrasts (Tone 2-Tone 3 was least discriminable) and that the discrimination rankings were the same for the AV and AO conditions, when stimuli were presented without noise. In CI speech where F0 is not available, pairings involving T3 were better discriminated, and this advantage was more pronounced in the AV condition. In the VO mode, Tone 2–Tone 3 was most easily discriminated while pairings involving Tone 4 were poorly discriminated. Finally, Burnham et al. (2015) reported for Cantonese that the dynamic Rising-Falling contrast was most discriminable in the VO mode and that static-dynamic pairs were better discriminated when they included the rising tone rather than the falling tone. Visual augmentation in noise was also greatest for the Rising-Falling contrast. Taken together, greater visual benefits are found for more dynamic tones or tone pairs that are more contrastive in contour shape. These results are consistent with findings in production that head movements are greater for tones with a larger amount of variance in F0 (Garg et al., 2019; Munhall et al., 2004; Yehia et al., 2002).

9.3.2 Non-native Perceivers

Burnham, Lau, Tam, and Schoknecht (2001b) followed up on the initial study by Burnham et al. (2001a) by exploring the perception of the same Cantonese stimuli by English (non-native, non-tonal) and Thai (non-native, tonal) perceivers. An AX discrimination task was used which again included the AV, AO, and VO modes. Stimuli were presented in the clear and in babble noise at an SNR of −0.6 dB. For both the English and Thai groups, performance in the VO mode was significantly better than chance. While the Thai perceivers showed no significant difference between AV and AO modes in the clear condition, visual augmentation was significant in the babble noise condition. English perceivers, in contrast to the Thai perceivers, did significantly better in the AO as compared to the AV mode. In the babble noise condition, English perceivers also did significantly better than chance for all three modes but there was no advantage for the AV over the AO condition. In sum, when presented with visual information only, both perceivers of a tonal and a non-tonal language were able to use this information to distinguish the tones of a language they did not know. In addition, for speech embedded in noise, tonal Thai perceivers showed increased accuracy when visual information was added to the auditory signal but English perceivers did not. Even though non-native English perceivers can pick up visual cues to tone as evidenced by their better-than-chance performance in the VO mode, they do not always seem capable of integrating this information with the auditory information.

The literature on visual augmentation in tone perception indicates that, for non-native perceivers, there appears to be a universal advantage of the AV mode over

the AO mode when stimuli are presented in noise. In their comprehensive study of Thai tone perception by native perceivers of Mandarin, Cantonese, Swedish, and English, Burnham et al. (2015) found that tone perception was consistently better in the AV than AO mode for all language groups. The one exception is the finding that English perceivers were equally accurate in AV and AO modes for speech presented in noise, which was attributed to a floor effect (Burnham et al., 2001b). Interestingly, naïve Dutch perceivers' identification of Mandarin tones was found to be better in the AV than AO mode even for stimuli presented in the clear (Han, Goudbeek, Mos, & Swerts, 2019). Overall, a visual benefit obtains for non-native perceivers whose native language is a tone language, as well as non-native perceivers without any prior exposure to a tone language.

Moreover, there is a language-specific aspect to the processing of visual cues to tone in that non-native perceivers whose native language is non-tonal benefit more from visual information than tonal perceivers. For example, English perceivers outperformed Mandarin perceivers in their discrimination of Mandarin tones in the VO mode (Smith & Burnham, 2012). They were also better than perceivers of other tone languages (Mandarin, Cantonese, Swedish) in discriminating Thai tones in VO (Burnham et al., 2015). In a comparison of congruent and incongruent AV information, English perceivers relied more on facial information while Mandarin perceivers relied almost exclusively on auditory information (Hannah et al. 2017). In sum, these studies indicate that facial cues for tone are more likely used by non-native perceivers who find themselves in a challenging non-native phonetic situation. However, non-natives' superior performance in the VO mode does not necessarily transfer to the AV mode; the English perceivers in Burnham et al. (2015) were poorer at AV integration than the non-native tone perceivers.

Taken together, non-native visual tone perception appears to involve language-specific aspects as a function of perceivers' linguistic experience, just as is the case in non-native auditory perception (e.g., Ingvalson & Wong, Chap. 2; Lee & Wiener, Chap. 1 of this volume). Further research explores these aspects by focusing on visual tone perception in different speech styles and by tracing learning trajectories through visual tone perception training.

9.3.2.1 Clear Speech

Research has shown that acoustic cues to segmental contrasts (e.g., vowels or fricatives) tend to be exaggerated in clear, hyperarticulated speech (Ferguson & Kewley-Port, 2007; Maniwa, Jongman, & Wade, 2009; Leung, Wang, Jongman, & Sereno, 2016). These clear speech tokens are in turn more intelligible than their casual speech counterparts (e.g., Ferguson & Kewley-Port, 2002; Maniwa, Jongman, & Wade, 2008). Additionally, visual cues are also more pronounced in clear speech (e.g., Kim, Sironic, & Davis, 2011; Tang et al., 2015). Kim et al. (2011) established that the greater articulatory movement observed in clear speech produced in noise contributed to greater intelligibility in the AV mode. Much less is known, however,

about the perception of clearly produced tones, particularly about whether hyperarticulated visual cues can enhance perception in a linguistically challenging non-native setting. In one of the few studies, Han et al. (2019) did not find an overall significant difference between the perception of clearly and casually produced Mandarin tones by Dutch perceivers. Inspection of the four speakers used in this study revealed that perceivers did perform significantly better on the clearly produced tokens produced by two speakers, but did significantly worse on one of the other two speakers. In addition, analysis of individual tones showed that perceivers identified Tones 2 and 4 more quickly in clear than casual productions. This is probably because contoured tones are hyperarticulated to a greater degree (cf. Kim & Davis, 2001). The lack of any effect of speech style on Tones 1 and 3 may be because Tone 1 involves minimal hyperarticulation, while Tone 3 is the easiest to distinguish in natural style already (Chen & Massaro, 2008; Mixdorff et al., 2005). These results indicate that for non-native perceivers, there may be some visual cues associated with clear speech production of tone. Moreover, these results are consistent with the articulatory findings of greater and more dynamic facial movements of contour than level tones (Garg et al., 2019), indicating that these visual cues are linguistically relevant and can aid non-native perception when enhanced.

9.3.2.2 Perceptual Training

Research has shown that non-native listeners with little or no experience with a tone language can improve their tone perception with a relatively brief training program (Wang, Spence, Jongman, & Sereno, 1999). Beginning American learners of Mandarin improved their accuracy in tone identification by 21% after eight sessions of high-variability training in which they identified which tone of a given pair they had heard and received feedback (Wang et al., 1999). Listeners were trained and tested in the AO mode only. Very few training studies have considered augmenting AO tone training with visual information. As mentioned earlier, Chen and Massaro (2008) included a training phase during which participants were instructed to pay attention to mouth, head/chin, and neck movements. Trainees received one 45 min training session during which they were first shown a 15 min video that illustrated various articulatory correlates of each tone, followed by about 10–20 practice trials with feedback. Training and testing all took place in the VO mode only. Results showed that training yielded a 15% increase in tone identification, from 33% accuracy before training to 48% after training. Most recently, Kasisopa, Antonios, Jongman, Sereno, and Burnham (2018) conducted a systematic comparison of Mandarin tone training in the AO and AV modes. Specifically, participants received either AO training or AV training and were then tested in both the AO and AV modes. Eight groups of participants, consisting of 6- and 8-year-old monolingual or bilingual children with either a tone or non-tone background participated: Thai monolingual, English monolingual, English-Thai bilingual, and English-Arabic bilingual. While the effect of training was minimal for 6-year-olds, 8-year-olds did show improvement as a result of training. In particular, results showed that tone language experience, either monolingual or

bilingual, is a strong predictor of learning unfamiliar tones. 8-year-old monolingual children improved with AV training but not with AO training, whereas 8-year-old bilingual children improved with AO training and to a lesser extent with AV training. These findings provide longitudinal data supporting linguistically motivated AV tone perception in that visual tone perception can be improved as a function of linguistic experience.

9.4 Gestural Tone Perception

Previous research has claimed that pitch is audio-spatial in representation (Connell et al., 2013). Findings from visual cues in tone production indicate that facial gestures such as eyebrows may be spatially equated to tonal variations. This association may encourage the perception of tone to be tied to visual-spatial features.

Indeed, similar to facial-spatial gestures, upward and downward hand gestures have been shown to affect pitch perception in the direction of the gesture (Connell et al., 2013). Capturing pitch in gestures owes its inspiration to the illustrative aids in musical perception. For example, to create audio-spatial connections, music teachers use hand gesture levels and diagrams of melodic contours to enhance pitch perception (Apfelstadt, 1988; Welch, 1985), and singers can be trained to improve their pitch perception accuracy using such gestures (Liao, 2008; Liao & Davidson, 2016). In a linguistic context, gestures have been shown to affect the perception of prosodic information in a non-native language. For example, Kelly, Bailey, and Hirata (2017) found that upward or downward hand movements congruous with the direction of the intonational pitch contour (rising or falling, respectively) could facilitate perception of intonation in a non-native language, while incongruous gesture-pitch matching was disruptive, suggesting a direct link between hand gestures and pitch perception.

The contribution of hand gestures has been shown to facilitate lexical tone perception and learning. Morett and Chang (2015) trained English perceivers to learn Mandarin tone words either with or without viewing hand gestures tracing tone contours. The results showed greater post-training improvements for the group who received training with gesture compared to the no-gesture training group. However, this improvement only held true for the word-meaning association task with a limited number of training words, whereas for tone identification, gestural training did not show any advantage. As such, it is not clear whether the facilitative effects of gesture could be attributed to effective cross-modal pitch-gesture associations or are the result of memorization using verbal labeling strategies (cf. Connell et al., 2013).

Hannah et al. (2017) attempted to address whether a pitch-gesture association can be established in a linguistically meaningful manner by manipulating auditory and gestural input and through comparing native and non-native perceptual patterns. Specifically, native Mandarin perceivers and native English perceivers identified Mandarin tones embedded in noise with either congruent or incongruent auditory-gestural inputs, where the hand movements tracing the tones were in the same (congruent) or different (incongruent) direction and shape as the tone contours. Native

Mandarin results showed exclusive reliance on auditory information in the congruent condition; whereas in the incongruent conditions, identification was partially based on gestures, demonstrating the use of gestures as valid cues in Mandarin tone identification. The English perceivers' performance improved significantly in the congruent auditory-gesture condition compared to the condition without gestural information. Moreover, they relied more on gestural than auditory information in the incongruent condition. These results reveal positive effects of gestural input (tracing tone contours) on both native and non-native tone perception, indicating that cross-modal (visual-spatial) resources can be recruited to aid linguistic perception. These patterns are consistent with the finding that visual presentation of schematic representations of the pitch contours enhances auditory tone perception (Liu et al., 2011, also see Ingvalson & Wong, Chap. 2 of this volume).

The fact that perceivers can establish a cross-modal link reflects a linguistically meaningful association between auditory and visual-spatial events in tone perception. Moreover, the different audio-gestural weighting patterns exhibited in native versus non-native perception further reveal the contribution of language-specific factors in multi-modal tone perception.

9.5 Concluding Remarks

In tone production, although head and eyebrow motion may be attention-drawing overall, the results of aligned head, eyebrow, and lip movements with specific spatial and temporal pitch movement trajectories of different tones suggest linguistically meaningful associations of these visual cues to tone articulation. Consistently, the degree of visual benefits in tone perception corresponds to the extent of contour movement dynamicity and shape contrastivity of individual tones. Moreover, these benefits can be particularly augmented in non-native perception, when the tonal signal is enhanced in clear speech and with the additional aid of hand gestures, and as perceivers gain additional experience in tone learning. These data suggest language-specific mechanisms and the influence of language experience in cross-modal tone production and perception, above and beyond a general language-universal system.

However, due to the scarcity of research in this area, there remain several under-explored research directions for a more thorough understanding of the role of visual cues in multi-modal tone production and perception. First, it is unclear how some visual cues (e.g., lips) are used in characterizing individual tones. Moreover, further evidence from tone perception research is needed to, further evidence from tone perception research is needed to determine how the visual tonal cues are used to facilitate perception of categorical tonal distinctions, and the extent to which perception is based on the linguistic relevance of these cues.

Understanding the visual correlates of tone production and perception will not only advance research on cross-modal integration of sensory-motor information in speech processing, but it will also have important applications for the development of effective tools for tone language acquisition and learning, as well as

audio-visual tonal speech synthesis including visual aids for impaired conditions and noisy environments.

References

Apfelstadt, H. (1988). What makes children sing well? *Applications of Research in Music Education, 7,* 27–32.

Attina, V., Gibert, G., Vatikiotis-Bateson, E., & Burnham, D. (2010). Production of Mandarin lexical tones: Auditory and visual components. In *Proceedings of International Conference on Auditory-visual Speech Processing (AVSP) 2010*, Hakone.

Burnham, D., Ciocca, V., & Stokes, S. (2001a). 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.

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001b). Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by nontonal language speakers. In D. Massaro, J. Light, & K. Geraci (eds.), *Proceedings of International Conference on Auditory-visual Speech Processing (AVSP) 2001*, Adelaide, SA, pp. 155–160.

Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., et al. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics, 36,* 1459–1491.

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., & Ciocca, V. (2006). The perception and production of phones and tones: The role of rigid and non-rigid face and head motion. In *Proceedings of the International Seminar on Speech Production 2006*, Ubatuba.

Campbell, R., Dodd, B., & Burnham, D. (1998). *Hearing by Eye II: Advances in the Psychology of Speechreading and Audio-visual Speech.* Hove, UK: Psychology 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.

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition, 81,* 124–130.

Desai, S., Stickney, G., & Zeng, F. G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *Journal of the Acoustical Society of America, 123,* 428–440.

Dohen, M., & Loevenbruck, H. (2005). Audiovisual production and perception of contrastive focus in French: A multispeaker study. *Interspeech, 2005,* 2413–2416.

Dohen, M., Loevenbruck, H., & Hill, H. (2006). Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variability. In R. Hoffmann & H. Mixdorff (eds.), *Speech Prosody 2006*, pp. 221–224.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 112,* 259–271.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research, 50,* 1241–1255.

Flecha-Garcia, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52,* 542–554.

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

Garg, S., Hamarneh, G., Jongman, Sereno, J.A., & Wang, Y. (2019). Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. *Speech Communication*, *113*, 47–62.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103,* 2677–2690.

Han, Y., Goudbeek, M., Mos, M., & Swerts, M. (2019). Effects of modality and speaking style on Mandarin tone identification by non-native listeners. *Phonetica, 76,* 263–286. https://doi.org/10.1159/000489174.

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology, 8,* 2051.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge: Cambridge University Press.

Huron, D., & Shanahan, D. (2013). Eyebrow movements and vocal pitch height: Evidence consistent with an ethological signal. *Journal of the Acoustical Society of America, 133,* 2947–2952.

Ishi, C. T., Haas, J., Wilbers, F. P., Ishiguro, H., & Hagita, N. (2007). Analysis of head motions and speech, and head motion control in an android. Paper presented at the *International Conference on Intelligent Robots and Systems*, San Diego, CA.

Jongman, A., Wang, Y., & Kim, B. (2003). Contribution of semantic and facial information to perception of non-sibilant fricatives. *Journal of Speech, Language & Hearing Research, 46,* 1367–1377.

Kasisopa, B., El-Khoury Antonios, L., Jongman, A., Sereno, J. A., & Burnham, D. (2018). Training children to perceive non-native lexical tones: Tone language background, bilingualism, and auditory-visual information. *Frontiers in Psychology, 9,* 1508. https://doi.org/10.3389/fpsyg.2018.01508.

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology 3*(7). https://doi.org/10.1525/collabra.76.

Kim, J., & Davis, C. (2001). Visible speech cues and auditory detection of spoken sentences: An effect of degree of correlation between acoustic and visual properties. In *International Conference on Auditory-visual Speech Processing (AVSP) 2001*, Aalborg.

Kim, J., & Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer, Speech and Language, 28,* 598–606.

Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication, 57,* 317–330.

Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception, 40,* 853–862.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT.

Leung, K., Jongman, A., Wang, Y., & Sereno, J. A. (2016). Acoustic characteristics of clearly spoken english tense and lax vowels. *Journal of the Acoustical Society of America, 140,* 45–58.

Liao, M. Y. (2008). The effects of gesture use on young children's pitch accuracy for singing tonal patterns. *International Journal of Music Education, 26,* 197–2113.

Liao, M. Y., & Davidson, J. W. (2016). The use of gesture techniques in children's singing. *International Journal of Music Education, 25,* 82–94.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning, 61,* 1119–1141.

Maniwa, K., Jongman, A., & Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *Journal of the Acoustical Society of America, 123,* 1114–1125.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken english fricatives. *Journal of the Acoustical Society of America, 125,* 3962–3973.

Mixdorff, H., Hu, Y., & Burnham, D. (2005). Visual cues in Mandarin tone perception. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, ISCA, Bonn, Germany, pp. 405–408.

Morett, L. M., & Chang, L.-Y. (2015). Emphasizing sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language and Cognitive Neuroscience, 30,* 347–353.

Moisik, S. R., Lin, H., & Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *Journal of the International Phonetic Association, 44,* 21–58.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15,* 133–137.

Perkell, J. S., Zandipour, M., Matthies, M. L., & Lane, H. (2002). Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *Journal of the Acoustical Society of America, 112,* 1627–1641.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Attention, Perception and Psychophysics, 77,* 571–591.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences, 104,* 7295–7300.

Scarbourough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech, 51,* 135–175.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences, 102,* 18748–18750.

Shaw, J. A., Chen, W. R., Proctor, M. I., Derrick, D., & Dakhoul, E. (2014). On the inter-dependence of tonal and vocalic production goals in Chinese. Paper presented at the *International Seminar on Speech Production (ISSP)*, Cologne, Germany.

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.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in english. *Language and Speech, 46,* 1–22.

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

Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics, 38,* 197–206.

Tang, L., Hannah, B., Jongman, Sereno, Wang, Y., & Hamarneh, G. (2015). Examining visible articulatory features in clear and plain speech. *Speech Communication*, *75*, 1–13.

Traunmüller & Öhrström. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*, 244–258.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106,* 3649–3658.

Welch, G. F. (1985). A schema theory of how children learn to sing in tune. *Psychology of Music, 13,* 3–18.

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

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