Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

The visual perception of lexical tone

Damien Smith (1), Virginie Attina (1), Connie So (1) and Denis Burnham (1)

(1) MARCS Auditory Laboratories, University of Western Sydney, Australia

PACS: 43.71.Hw 43.71.Es 43.71.Rt

ABSTRACT

Visual speech (lipreading) supports speech perception not only when the auditory signal is limited, degraded, or miss-ing, but also in mismatched auditory and visual speech component conditions when the auditory signal is clear and undegraded. While most of this research has been done with segments (consonants and vowels), research at MARCS has now provided evidence for visual speech cues for non-segmental features of language, particularly lexical tone. Early work has shown that Cantonese adults identify visual-only words differing only on tone at a rate significantly above chance; that even non-Cantonese tone (Thai) and non-tone (English) language speakers use visual information to discriminate words differing only in tone; and that one of the most likely vehicles for visual tone information is minute rigid movements of the head. In a more recent study, we investigated visual augmentation for discrimination of Mandarin tones, with F0 information degraded by using simulated cochlear implant (CI) audio. Native Mandarin and Australian English speakers were asked to discriminate between minimal pairs of Mandarin tones in five condi-tions: Auditory-Only, Auditory-Visual, CI-simulated Auditory-only, CI-simulated Auditory-Visual, and Visual-Only. Discrimination in CI-simulated audio conditions was poor compared with normal audio, but the availability of visual speech information improved discrimination in CI-simulated audio conditions, particularly on tone pairs with strong durational cues, but also for some pairs cued primarily by F0 cues. In Visual-Only, both Mandarin and Australian English speakers discriminated tones above chance, and interestingly, tone-naïve listeners outperformed native listen-ers, suggesting firstly that visual speech information for tone is available and may be under-used by normal-hearing tone language perceivers, and secondly that the perception of such information may be language general, rather than the product of language specific learning. In a follow-up study with English-language children, it was found that point-light reductions of visual tone information did not augment tone perception, but tone perception was stronger when their pitch contours were presented as violin sounds rather than as natural speech. In future studies along this line, groups of Cantonese, Thai and Australian English children (4 to 10 years old) and adults will be tested in a dis-crimination task using auditory-only and audio-visual Cantonese and English stimuli to see how visual information (coming from both consonants and vowels or tones) is used across ages and languages.

INTRODUCTION

Speech perception is not solely an auditory phenomenon. Use of visual (facial) speech information by the hearing impaired is well-documented e.g.[1], and those with normal hearing have been shown to use visual speech information when the auditory signal is unavailable or degraded [2-4]. Possibly the most dramatic demonstration of this is the “McGurk Effect” [5] where the syllable /ba/ is presented auditorily, simultaneously with a visual /ga/. The incongruent auditory and visual speech syllables are integrated into one speech percept – not /ba/, nor /ga/, but an intermediate syllable such as /ða/ or /da/. This effect is robust, occurring with undegraded auditory and visual recordings, and is persistent even when the perceiver is entirely aware of the illusion. An important feature of the McGurk effect is that multimodal speech perception does not involve “capture” of one modality by another. Instead of visual speech clarifying incomplete auditory information, here visual speech information actively contributes to the percept, despite being in direct conflict with a clear auditory signal.

A striking feature of auditory-visual speech perception is the variation in use both cross-linguistically and among speakers of the same language. Extensive work using a variant of the McGurk paradigm by Massaro and colleagues [6-8] points strongly to large individual differences in visual perception of speech, which are in turn linked to differences in how incongruent McGurk stimuli are integrated. Likewise, Smith and Croot [9] found considerable individual differences in visual augmentation in noise that predicted English-speaking participants' later discrimination of a non-native (Italian) length distinction using visual information. Importantly, in each of these studies, participant's ability to use visual speech was found to be independent of their auditory perception of speech. The source of this variation may be raw psychophysical attributes – speechreading ability among deaf perceivers has been linked to visual movement sensitivity [1], although this relationship was not observed among hearing perceivers. Another factor may be language background – Japanese-speaking perceivers have been found to show less auditory-visual integration than English-speaking perceivers [10-13], and Japanese-speaking children do not show a developmentally-linked increase in auditory-visual speech integration, while English-speaking children do [11]. Compared with English, Japanese has a greatly restricted vowel space and far fewer complex consonant clusters - visual speech may simply be less useful in this environment. Another possibility is that it is linked to cultural differences, in that it is considered impolite in Japan to look directly into the face of the person talking to you. Both of these explanations depend

on the thesis that ability to use visual speech information depends on prior experience in using it. Supporting this, a case study on an "expert" speechreader suggests that optimal speechreading ability may depend on a combination of high working memory capacity, excellent phonological skills, and extensive practice [14].

It would seem intuitively that visual speech information should be of most use in perception of segmental components of speech, and indeed there are consistent findings to suggest that consonants, particularly those articulated nearer the front of the mouth (such as bilabials) are more readily perceived via visual speech e.g. [15]. However, there is a growing body of evidence suggesting there are visual speech cues for suprasegmental features of speech, such as intonation, stress, and lexical tone.

Lexical tone refers to variation in pitch height and contour that alter meaning on a word level. Over 70% of the world's languages are tonal [16] and roughly 50% of the world's population speak tonal languages [17]. While the core cue for lexical tone is F0, in many languages there are often additional auditory cues such as vowel duration, amplitude envelope, and voice quality that appear consistently with tones [18-20].

Burnham, Ciocca and Stokes [21] found that Cantonese speakers were able to identify Cantonese tones slightly above chance using visual footage only in some conditions: in running speech (as opposed to citation form); on monophthongs (as opposed to diphthongs); and on contour tones (as opposed to level tones). A stronger effect was found by Burnham, Lau, Tam and Schoknecht [22], investigating Thai- (a tonal language) and English-speaking listeners discrimination of Cantonese tones presented in auditory-visual (AV), audio-only (AO) and visual-only modalities (VO). Both English- and Thai-speaking perceivers were able to discriminate tones above chance levels in the VO condition (as well as in the auditory conditions), and with noise added, Thai listeners displayed visual augmentation (improved discrimination in the AV condition compared with AO). These findings suggest the presence of visual information for tone that is available cross-linguistically, both to non-tonal language speakers and to speakers of other tonal languages.

A small effect of visual augmentation in noise (but not in clear audio) was also found for Mandarin speakers' identification of Mandarin tones [23], although this effect was not found with F0 synthetically devoiced. Similar findings were also made for Thai [24] and Vietnamese tones [25].

The nature of visual cues for tone is not entirely clear, but there is some evidence to suggest that head motion plays a role. Rigid movements of the head have been linked to perception of Cantonese tones [26], and this in turn is consistent with links between head movements and perception of prosody [27-29]. While there is no direct anatomical reason for this connection, a suggested mechanism is that tilting the head back puts a small strain on the cricothyroid cartilage, which in turn slightly tightens the vocal folds. While this relationship would not be enough to constrain (or contribute importantly) to spoken pitch, it may lead to consistent head motion in line with spoken F0 (and hence with prosody and with lexical tone) [30]. This is also supported by prosodic discrimination based solely on visual presentation of the top part of the head [31]. Other possible visual correlates of pitch include eyebrow movements Cavé et al., [32], or laryngeal movements.

Chen and Massaro [33] found native Mandarin speakers' identification of Mandarin tones in isolation slightly above chance using visual cues only. The experimenters observed some consistent patterns of head movements and movements in the neck area for each talker, and subsequently instructed participants to attend to these features, resulting in substantially improved tone identification.

Thus it is appears that there are visual cues for lexical tone. However, the nature of these cues is not entirely clear, and in the light of the result from Chen and Massaro [33], it appears they may not be fully utilised by all listeners.

EXPERIMENT 1

In Experiment 1, we investigated discrimination of Mandarin lexical tones by native and tone-naïve (Australian English speaking) listeners in five modalities: Visual-only (silent) (VO), Auditory-Only (AO), Auditory-Visual (AV), Auditory-Only with simulated cochlear implant audio (AO-C) and Auditory-Visual with simulated cochlear implant audio (AV-C). Cochlear implants provide direct electrical stimulation to the cochlea, dividing the auditory signal to set of frequency-centred bands and delivering a pulse code accordingly using an array of up to 22 electrodes, taking advantage of the tonotopic arrangement of the basilar membrane. As such, cochlear implants have had great success in allowing the profoundly deaf to perceive segmental components of speech; however, the dramatic reduction in auditory resolution (compared to the roughly 6000 inner hair cells of the cochlea) results in users having great difficulty interpreting pitch information in music [34-37], prosody, and lexical tone [38-40]. Cochlear implant percepts can be approximately simulated by resynthesising the pulse code from a cochlear implant processor into an auditory signal, using a sine or filtered noise carrier signal [41]. This procedure was taken advantage of to investigate the availability of tone information in speech that, while recognisable as speech, does not contain any readily interpretable F0 cues for tone.

METHOD

Participants

Participants were 48 native speakers of Mandarin Chinese (27 female, 21 male, M_{age}=23), and 48 Australian English speaking monolinguals (34 female, 14 male, M_{age}=19), with no reported exposure to a tone language. Mandarin speakers were recruited via advertisement at the University of New South Wales and reimbursed $20 for travel expenses, while Australian English speakers were first year psychology students at the University of Western Sydney, participating for course credit.

Stimuli

36 Mandarin monosyllables were selected for use in this experiment, comprising minimal tone pairs across the four Mandarin tones (making six tonal contrasts) in three broad phonetic groups: consonant-vowel, glide-vowel, and consonant-glide-vowel, balanced across pairs for lexical frequency as closely as possible. Auditory-visual recordings were made of two native speakers of Mandarin (1 male, 1 female) producing eight repetitions of each word. Speakers were recorded directly facing a digital video camera 1.5m away and saved in MPEG2 format. Audio was recorded using a separate microphone and saved in 48 kHz, 16 bit .wav format. Utterances were each cut into a separate video file, with the first and last repetition discarded, making for a total of 6 repetitions x 36 words x 2 speakers = 432 tokens. Auditory-only and visual-only stimuli were generated by isolating the audio and video from the recordings, respectively.

For the simulated cochlear implant audio, tokens were processed using the Nucleus MATLAB Toolbox 4.20 [42], converting the audio into a pulse stream using an Advanced Combinatorial Encoding (ACE) map [43], choosing 12 of 22 channels at 6000 pps. The resultant pulse streams were then resynthesised using a pink noise carrier and saved as separate .wav files. These were then dubbed onto their original video files to create the AV-C stimuli.

Procedure

Each participant completed 480 AX trials, with an ISI of 500ms, comprising 96 trials each in the 5 experimental conditions (VO, AO, AV, AO-C, AV-C). Stimuli were presented to participants via laptop screen and headphones. Participants were asked to respond via keypress to indicate for each trial whether the two words presented had the "same" or "different" tone.

RESULTS AND DISCUSSION

In the VO condition, it was found that both Mandarin-speaking and English-speaking participants discriminated tones slightly above chance ($t(95) = 10.9$, $p<.001$; see Figure 1). Strikingly, the non-tonal language listeners displayed better discrimination of tone than the native listeners ($F(1,94) = 7.009$, $p <.01$), suggesting firstly that visual cues for tone may be language general, rather than available only to experienced listeners, and secondly suggesting that while visual cues for tone may be available, they may not be entirely taken advantage of by native listeners.

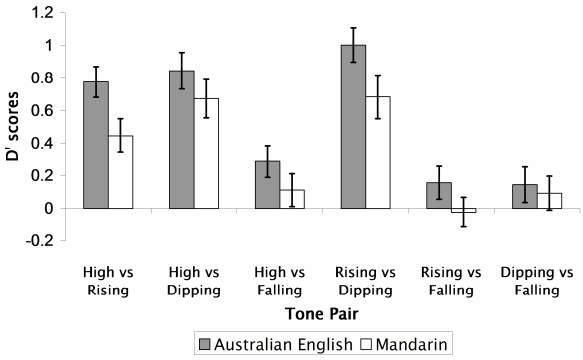

Figure 1. D' scores from the visual-only condition, for Australian English monolinguals and native Mandarin speakers.

Some consistent patterns emerged from the VO data. Mandarin has a rich tone space, comprising 4 tones: 1st tone ("High"), with a steady, high pitch; 2nd ("Rising"), which dips briefly below a mid level before rising to a high level; 3rd ("Dipping", "Low rising", or "falling-rising"), which dips lower than the 2nd tone before rising more slowly, and 4th ("Falling"), which starts high and rapidly drops to a low level. In this experiment, the tone contrasts most easily discriminated were the Rising vs Dipping contrast, the High vs Dipping contrast and the High vs Rising contrast. The High vs Falling contrast was also slightly more easily discriminated than Rising vs Falling and Dipping vs Falling. Importantly, it seems that contrasts involving a level (High) tone and a contour tone are the most readily discriminated visually, with the exception of Rising vs Dipping. In the case of Rising vs Dipping, the tones have a similar contour but display a marked difference in duration. Even among native speakers, Rising vs Dipping is the most easily confused contrast auditorily [44], which was also reflected in the AO data in this study. It may be that the absence of the rising intonation cue common to both tones allowed listeners to attend solely to a clearly visible difference in duration.

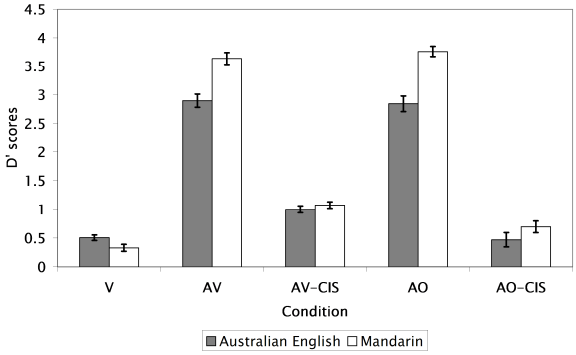

Figure 2. D' scores across all five experimental conditions

As expected, both native and non-native listeners showed slight visual augmentation for tone discrimination in clear audio (AV compared with AO, see figure 2). Also as expected, discrimination in the CI audio conditions was markedly worse for all participants than in clear audio, but much stronger visual augmentation was observed in these conditions.

These data are consistent with the general premise that there exist visual cues for lexical tone. The stronger effect for level vs. contour contrasts may provide support for the rigid head motion account of visual tone – level tones would be stationary, while contour tones would involve movement. If this were the case, however, one might expect much stronger discrimination for Rising vs. Falling and Dipping vs. Falling. However, all three of the tones involved in this contrast begin with a falling pitch, leading to a visual correlate that is difficult to distinguish. Support for this comes from data from the AV-C and AO-C conditions – Rising vs Falling and Dipping vs Falling show large Auditory-Visual augmentation, despite being barely discriminable in VO. In this case, it may be that while the auditory signal does not carry F0 information, it could serve to disambiguate whether an upwards head motion is part of the vowel (as presumed for the upward rise in the Rising and Dipping tones) or simply the head returning to initial position (as presumed for the Falling tone. To establish this it will be necessary to analyse the videos used in this experiment for these cues.

A larger question is why non-native speakers were better able to discriminate Mandarin tones in silence than native speakers. We offer three possible explanations. The first is that listeners have been shown to attend more to visual speech when listening to a non-native speaker [45],[46],[10],[11],[13] – it may be that the Australian English speaking listeners treated Chinese faces speaking Chinese words as "non-native", whereas for the Mandarin-speaking listeners, both the faces and the words were "native".

A second possibility is that visual information might normally be ignored as a cue for tone by Mandarin speakers, due to its comparatively low utility. Dutch speakers have been shown to more readily identify stressed and unstressed syllables of English words in isolation than native English speakers [47]. This may be due to stress being a useful cue for word boundaries in Dutch, but not in English, and hence it is ignored (even though it is consistently produced) by English speakers. In an analogous fashion, Mandarin speakers may simply be ignoring the less consistently available visual cues for tone.

A third possibility is that Chinese listeners attend less to visual speech in general. Chinese listeners show weaker McGurk effects than Japanese listeners [48], who in turn show weaker McGurk effects than English speaking listeners [11],[13]. This may once again be a result of a less visually differentiable phonological space in Mandarin than in English – while Mandarin has a wide range of fricative and affricate contrasts and a vowel space similar to English [49], Mandarin does not have the complex consonant clusters in initial, medial and final position that riddle English speech.

In a second study, we took quite a different approach, investigating how linguistic experience affects the perception of non-native tones. Perception of Cantonese tones by native Australian English-speaking children in three different age groups (4-, 6-, and 8-year-old) was examined and compared to that of Australian English adults and Hong Kong Cantonese adults across four modalities: Auditory Only (AO), Auditory-visual (AV), Non-Speech Auditory Only (AO-N) and Non-Speech Auditory-visual (AV-N).

EXPERIMENT 2

METHOD

Participants

Three groups of Australian English monolingual children participated: 18 four-year-old children, 18 six-year-old children, and 20 eight-year-old children. Children were recruited from the MARCS Babylab register, and their caregivers received a gift bag and $30 travel reimbursement. Adult participants comprised 20 Australian English monolinguals and 18 Hong Kong Cantonese-speaking adults, recruited via word of mouth in the Western Sydney area.

Stimuli

Auditory speech stimuli were created from recordings of a female native Cantonese speaker's tone productions on Cantonese monosyllables. The visual component was an animated human face, constructed based on the scanned facial posture data of real human faces via principal component analysis and driven by Optotrak [50] motion capture data. Non-speech stimuli were either Auditory-only or auditory-visual recordings of a musician imitating Cantonese tones with a violin.

Procedure

Perception was investigated using a Go/No-Go perceptual discrimination task, in which participants were played one token repeatedly, and were asked to respond (via button press) when the token changed to to one of another class. The advantage of using this method is that it can be used to assess perception in children as well as in adults. The experimental session for each participant comprised eight blocks of Go/No-Go trials, two of each of the four modalities, with tone pairs counterbalanced across participants. In all conditions, tones were presented with either white noise or babble, counterbalanced across participants. This served an analogous function to the CI audio presentation in Experiment 1, increasing perceptual difficulty in the hope of teasing out evidence of visual augmentation.

RESULTS AND DISCUSSION

Accuracy scores from the Australian English speaking adults and children were analysed in a 4 x (2 x 2) ANOVA, with age between subjects and context (Auditory-Only or Auditory-Visual) and mode (Speech or Non-speech) within subjects. No visual augmentation was observed in this experiment, with no significant difference between auditory-only and auditory-visual in either children or adults, speech or non-speech (see Figure 3). Tone perception was improved with age, with Australian-English adults outperforming the three groups of children ($F(1, 72) = 19.52$, $p < .001$), and 8-year-olds outperforming the 6- and 4-year-olds across all conditions ($F(1, 72) = 24.27$, $p < .001$).

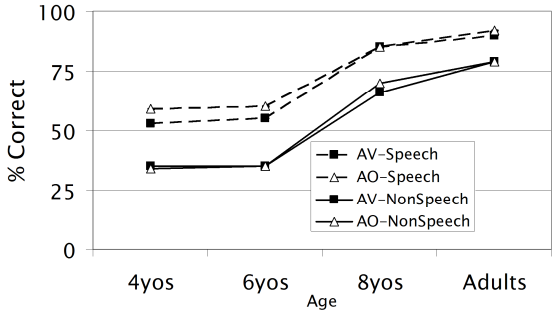

Figure 3. Percentage correct responses for Australian English monolingual children and adults.

Data from the native Cantonese and Australian English speaking adults was analysed in a 2 x (2 x 2) ANOVA, with language background between subjects and context and modality within subjects. No main effects were found of langauge background, or of context, but there was a main effect of mode, with improved perception in the Non-Speech condition for both Cantonese and Australian-English speaking adults ($F(1, 36) = 7.50$, $p < .01$) (Figure 4).

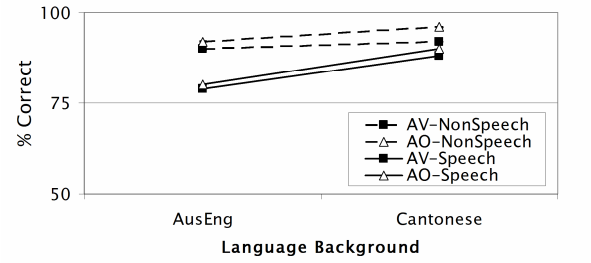

Figure 4. Percentage correct responses for Australian English monolingual adults and Native Cantonese speaking adults.

Visual speech integration has been shown previously to develop later than many other speech perception skills, but is expected to emerge for segmental information around the age of 7-8 [11]. It may be that visual speech information for tones does not follow this developmental course, explaining the child results. As for adults, it may simply be that discrimination in clear audio in a Go/No-Go task is too easy a task to pick out subtle visual augmentation effects in adults. Accordingly, an expanded version of this experiment is proposed, across children from 4 to 10 years old, from Thai, Cantonese and Australian English backgrounds, and on visual perception of tones, vowels and consonants, in order to tease out developmental, psycholinguistic and cultural factors in the visual perception of tone. In addition to this, further experiments on adult native Australian English and Mandarin speakers are proposed using eyetracking to investigate both differences between native and non-native looking patterns in a visual tone task, and equally to explore whether a consistent strategy emerges among successful visual tone perception.

REFERENCES

1 T. Mohammed, R. Campbell, M. Macsweeney, E. Milne, P.

Hansen, and M. Coleman, "Speechreading skill and visual movement sensitivity are related in deaf speechreaders," *Perception*, **34**, 205-216, (2005).
2 L.E. Bernstein, E.T. Auer, and S. Takaganayagi, "Auditory speech detection in noise enhanced by lipreading," *Speech communication*, **44**, 5-18, (2004).
3 N.P. Erber, "Discussion: Lipreading Skills," *Sensory Capabilities of Hearing Impaired Children*, ed.R.E. Stark, (Baltimore, University Press, 1974), pp. 69-73.
4 W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. America*, **26**, 212-215, (1954).
5 H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, **264**, 746-748, (1976).
6 D.W. Massaro, *Perceived Talking Faces: From Speech Perception to a Behavioral Principle*, (Cambridge, MA, MIT Press, 1998).
7 D.W. Massaro, "Speech perception," *International Encyclopaedia of Social and Behavioral Sciences*, eds.N.M. Smelser, P.B. Baltes, and W. Kintsch, (Amsterdam, Elsevier, 2001), pp. 14870-14875.
8 D.W. Massaro and D.G. Stork, "Speech Recognition and Sensory Integration," *American Science*, **86**, 236-244, (1998).
9 D.J. Smith and K. Croot, "Visual contributions to learning an unfamiliar language contrast," 33rd Australian Experimental Psychology Conference, Brisbane, Australia. In *Aust. J. Psych*, **58** (suppl. 1), 5, (2006)
10 P.K. Kuhl, M. Tsuzaki, Y. Tohkura, and A.N. Meltzoff, "Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces," *Third International Conference on Spoken Language Processing*, 1994.
11 K. Sekiyama and D. Burnham, "Impact of language on development of auditory-visual speech perception," *Developmental Science*, **11**, 306, (2008).
12 K. Sekiyama and Y. Tohkura, "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility.," *J. Acoust. Soc. Am*, **90**, 1797-1805, (1991).
13 K. Sekiyama and Y. Tohkura, "Inter-language differences in the influence of visual cues in speech perception.," *J.Phon.*, **21**, 427–444, (1993).
14 J. Rönnberg, J. Andersson, S. Samuelsson, B. Söderfeldt, B. Lyxell, and J. Risberg, "A Speechreading Expert: The case of MM," *J. Speech Lang. Hear. R.*, **42**, 5-20, (1999).
15 K. Nielsen, "Segmental Differences in the visual contribution to speech intelligibility.," *UCLA Working Papers in Phonetics*, **103**, 106-147, (2004).
16 M.J.W. Yip, *Tone*, (Cambridge, Cambridge University Press, 2002).
17 V.A. Fromkin, *Tone: A Linguistic Survey*, (New York, Academic Press, 1978).
18 Q.J. Fu and F.G. Zeng, "Identification of temporal cues in Chinese tone recognition," *Asia Pacifica Journal of Speech, Language and Hearing*, **5**, 45-57, (2000).
19 Q.J. Fu, F.G. Zeng, R.V. Shannon, and S.D. Soli, "Importance of tonal envelope cues in chinese speech recognition," *J. Acoust. Soc. Am*, **104**, 505-510, (1998).
20 D.H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, **49**, 25-47, (1992).
21 D. Burnham, V. Ciocca, and S. Stokes, "Auditory-Visual Perception of Lexical Tone," *Eurospeech 2001*, (Scandinavia, 2001).
22 D. Burnham, S. Lau, H. Tam, and C. Schoknecht, "Visual discrimination of cantonese tone by tonal but non-cantonese speakers, and by non-tonal language speakers," *avsp 2001*, 2001).
23 H. Mixdorff, Y. Hu, and D. Burnham, "Visual cues in Mandarin Tone perception," *Interspeech 2005*, (Lisbon, Portugal, 2005), pp. 405-408.
24 H. Mixdorff, P. Charnvivit, and D. Burnham, "Auditory-Visual Perception of Syllabic Tones in Thai," *AVSP-2005*, (British Columbia, Canada, 2005), pp. 3-8.
25 H. Mixdorff, C. Luong, D. Nguyen, and D. Burnham, "Syllabic Tone Perception in Vietnamese," *Proceedings of TAL 2006*, (La Rochelle, France, 2006), pp. 137-142.
26 D. Burnham, J. Reynolds, E. Vatikiotis-Bateson, H. Yehia, V. Ciocca, R. Haszard-Morris, H. Hill, G. Vignali, S. Bollwerk, H. Tam, and C. Jones, "The perception and production of phones and tones: The role of rigid and non-rigid face and head motion," *Proceedings of ISSP 2006, 7th International Seminar on Speech Production*, , eds.H. Yehia, D. Demolin, and R. Laboissiere, 2006), pp. 185-192.
27 E. Vatikiotis-Bateson and H. Yehia, "Physiological Modeling of facial motion during speech," *Trans. Tech. Comm. Psychol. Physiol. Acoustics*, **H-96-65**, 1-8, (1996).
28 H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," *The 5th International Seminar on Speech Production*, 2000), pp. 265-268.
29 H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *J. Phon.*, **30**, 555-568, (2002).
30 K. Honda, "Physiological factors causing tonal characteristics of speech: from global to local prosody," *Speech Prosody 2004*, 2004).
31 E. Cvejic, J. Kim, and C. Davis, "Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion," *Speech Communication*, (2010).
32 C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and F0 variations," *Proceedings of the ICSLP*, (Philadelphia, 1996), pp. 2175-2179.
33 T.H. Chen and D.W. Massaro, "Seeing pitch: Visual information for lexical tones of Mandarin-Chinese," *J. Acoust. Soc. Am*, **123**, 2356, (2008).
34 K. Gfeller, C. Turner, M. Mehr, G. Woodworth, R. Fearn, J. Knutson, S. Witt, and J. Stordahl, "Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults," *Cochlear Implants International*, **3**, 29-53, (2002).
35 V. Looi, H. McDermott, C. McKay, and L. Hickson, "Pitch discrimination and melody recognition by cochlear implant users," *International Congress series*, **1273**, 197-200, (2004).
36 H. McDermott and V. Looi, "Perception of complex signals, including musical sounds, with cochlear implants," *International Congress series*, **1273**, 201-204, (2004).
37 B. Swanson, P. Dawson, and H. Mcdermott, "Investigating cochlear implant place-pitch perception with the Modified Melodies test," *Cochlear Implants International*, **10**, 100-104, (2009).
38 D.K.K. Au, "Effects of stimuation rates on Cantonese lexical tone perception by cochlear implant users in Hong Kong," *Clinical Otolaryngology*, **28**, 533-538, (2003).
39 T.S. Huang, N.M. Wang, and S.Y. Liu, "Tone perception of Mandarin speaking postlingually deaf implantees using the Nucleus 22-channel cochlear implant mini system," *International cochlear implant, speech and hearing symposium*, (Melbourne, Annals of Otology, Rhinology and Laryngology, 1995).
40 K.Y.S. Lee, S.N. Chiu, and C.A. van Hasselt, "Tone Perception Ability of Cantonese-Speaking Children," *Language and Speech*, **45**, 387-406, (2002).
41 R. Burkholder, D. Pisoni, and M. Svirsky, "Perceptual learning and nonword repetition using a cochlear implant simulation," *International Congress series*, **1273**, 208-211, (2004).
42 B. Swanson and H. Mauch, *Nucleus Matlab Toolbox (Version 4.20)*, Sydney, Australia, Cochlear Corporation.
43 A.E. Vandali, L.A. Whitford, K.L. Plant, and G.M. Clark, "Speech perception as a function of electrical stimulation rate: using the nucleus 24 cochlear implant system," *Ear & Hearing*, **21**, 608-624, (2000).
44 D.L. Blicher, R.L. Diehl, and L.B. Cohen, "Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement.," *J. Phon.*, **18**, 37-49, (1990).
45 B. de Gelder, P. Bertelson, J. Vroomen, and H. Chen, "Interlanguage differences in the McGurk effect for Dutch and Cantonese listeners," *Proc. 4th European Conference on Speech Communication and Technology*, 1995).
46 H. Grassegger, "McGurk Effect in German and Hungarian listeners," *Proc. Int. Congr. Phonetic Sciences*, **3**, 210-213, (1995).
47 N. Cooper, A. Cutler, and R. Wales, "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners," *Language and speech*, **45**, 207, (2002).
48 Y. Hayashi and K. Sekiyama, "Native-Foreign Language Effect in the McGurk Effect: A Test with Chinese and Japanese," *AVSP-1998*, 1998), pp. 61-66.
49 S. Duanmu, *Phonology of Standard Chinese*, Oxford University Press: Oxford, 2000).
50 R.A. States and E. Pappas, "Precision and repeatability of the Optotrak 3020 motion measurement system," *J Med. Eng. Tech.*, **30**, 11, (2006).