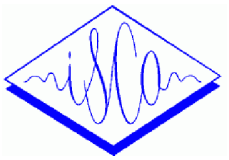
ISCA Archive
http://www.isca-speech.org/archive



FAAVSP - The 1st Joint Conference on
Facial Analysis, Animation, and
Auditory-Visual Speech Processing
Vienna, Austria,
September 11-13, 2015

Audio-Visual Perception of Mandarin Lexical Tones in AX Same-Different Judgment Task

Rui Wang[1], *Biao Zeng*[1], *Simon Thompson*[1]

[1] Psychology Department, Faculty of Science and Technology, Bournemouth University, UK
Rui.Wang@bournemouth.ac.uk

Abstract

Two same-different discrimination tasks were conducted to test whether Mandarin and English native speakers use visual cues to facilitate Mandarin lexical tone perception. In the experiments, the stimuli were presented in 3 modes: audio-only (AO), audio-video (AV) and video-only (VO) under the clear and two levels of signal-to-noise ratio (SNR) -6dB and -9dB noise condition. If the speakers' perception of AV is better than that of AO, the extra visual information of lexical tones contributes tone perception. In Experiment 1 and 2, we found that Mandarin speakers had no visual augmentation under clear and noise conditions. For English speakers, on the other hand, extra visual information hindered their tone perception (visual reduction) under SNR -9dB noise. This suggests that English speakers rely more on auditory information to perceive lexical tones. Tone pairs analysis in both experiments found that visual reduction in tone pair T2-T3 and visual augmentation in tone pair T3-T4. It indicates that acoustic tone features (e.g. duration, contour) can be seen and be involved in the process of audiovisual perception. Visual cues facilitate or inhibit tone perception depends on whether the presented visual features of the tone pairs are distinctively recognised or highly confusing to each other.

Index Terms: audiovisual speech perception, Mandarin, lexical tone

1. Introduction

Do you find talking over the phone is less intelligible than talking face-to-face? This would probably be worse when the phone signal is poor, but not for a communication in the noisy place when you can watch the speaker's face. Speech perception relies not only on auditory information but also on visual clues. Classic studies provided strong evidence that visual speech can facilitates or even distorts the speech perception. Sumby and Pollak [9] demonstrated that seeing the face contributed greater in low SNR condition. McGurk effect [6] is more robust example that visual information can interfere in speech perception. When the lip movement *ga* is presented with the audio syllable *ba*, it would be perceived as *da/tha*.

Many studies have had an insight look into consonants, only a few studies had tried to explore the audio-visual aspect of lexical tones. A possible reason may be that lexical tones are difficult to be seen from the face. Summerfield [10] explained that vision of a face compensates spectral cues in adverse condition, because visual cues supply the place of articulation (e.g. labial dental [f] / [v]) which is easily masked by the noise, but tone articulation relies on laryngeal source [5] which cannot be reflected visually on the face [4, 11], consequently visual cues might contribute little to tone perception. However, some studies [1, 2, 3, 7, 8] have shown some evidence that tonal language speakers can use visual information to facilitate their perception in the adverse condition. In this study, we are looking at whether visual cues will play a role in compensating Mandarin tone perception under the clear and noise condition by comparing to a few relevant studies in terms of the methodology and findings.

Burnham et al [2] employed a same-different discrimination task to test the Cantonese tones audio-visual perception by Thai native speakers and found the visual augmentation effect (audio-video perception is better than audio alone) under the noise. Furthermore, Burnham et al [3] used another same-different discrimination task to test Thai and Mandarin native speakers with Thai tones and they found that the augmentation was particularly greater on contour-contour tone pairs (231, 315) [1] rather than level-level/level-contour [2] tone pairs. Methodologically, in Burnham's studies, the presenting stimuli came from the same person in each experiment, which may cause bias that participants compare 2 face motions rather than speech-specific visual cues on the face. Although contour-contour tones contrast attained a clear augmentation in their study, there is no further explanation and discussion for what visual features of the contour tones contribute to visual augmentation. Despite that Burnham' studies observed tone perception (Thai) with tonal speakers (Mandarin speakers), they had not tested Mandarin speakers with Mandarin tones.

Mixdroff et al [8] have conducted an identification task which examined Mandarin native speakers' audiovisual tone perception with Mandarin tones. They reported that the perception of audio-video mode is better than that of auditory alone at the lower SNR level noise (-9dB, -12dB) and tone 3 is the most visually identifiable tone under the strong noise. The study explained the reason for that is visual tone 3 provides distinctive F0 cues – duration and intensity – compared to other tones. Based on this implication, if tone 3 can be easily visualised, then the tone pairs consisting of tone 3 may have greater visual augmentation effect (e.g. T1-T3, T2-T3, T3-T4). Given that contour-contour tone pair in Thai were the most visually detectable, Mandarin contour-contour pair containing tone 3 (T2-T3, T3-T4), hypothetically, would be easier to be discriminated.

In our study, we will cover the limitations mentioned above and test Mandarin audio-visual tone perception by employing 2 tone same-different discrimination (AX) tasks. In order to avoid the strategies that participants may use during the task,

the two tokens within individual trials will not be the same in mode (Experiment 1) or not be from the same speaker (Experiment 2). Two hypotheses will be tested in the following 2 experiments: 1) The visual augmentation effect appears under the noise condition; 2) The visual augmentation effect will be stronger on Mandarin contour-contour tones containing tone 3 (T2-T3, T3-T4) rather than the other tone pairs.

2. Experiment 1

2.1 Methods

2.1. 1. Participants and Materials

Twenty Mandarin native speakers (age: 22-30; female: 8) and 20 English native speakers (age: 19-30; female: 14) participated in the study. None of them had known hearing problems and their vision was normal or corrected to normal. All participants received a payment as a reward.

The experiment employed a monosyllable *bai* with 4 tones as a stimuli presented in 3 modes: audio-only (AO), audio-video (AV) and video-only (VO). Two levels of babble noise synthesized by 6 native Mandarin speakers were embedded in the stimuli, SNR-6dB and -9dB. The video materials were recorded by a male Mandarin native speaker in a noise-cancelled booth. The man in the video was only presented from the top of the head to the neck. The video clips were edited via Adobe Premiere and audio waveforms by Audacity. All stimuli were RMS normalised at -12dB.

Table 1. *Acoustic features of stimuli (bai)*

	bai1	bai2	bai3	bai4
Average Pitch (Hz)	123.88	116.43	95.48	118.22
Duration (s)	0.85	0.78	0.94	0.47
Intensity (dB)	66.28	68.86	62.27	64.78

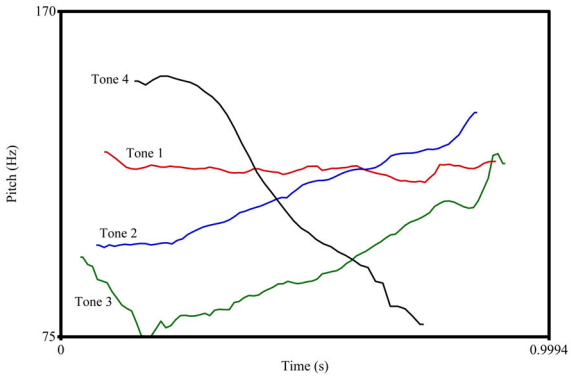


Figure 1. *Contours of 4 lexical tones of syllable (bai)*

2.1.2. Procedures

This experiment employed AX task that participants were told to judge whether two given syllables in succession were the same or different in terms of tones in each trial. AO, AV and VO stimuli were presented in three blocks respectively which were counterbalanced for every participant. Within each trial, the first token was presented in audio alone but the second one was either presented as AO, AV or VO mode in the corresponding block. The ISI between the two was 500 milliseconds (ms). In order to induce the visual augmentation effect, two levels of (SNR -6dB, -9dB) babble noises were embedded in the second token. Together with the clear condition, 3 levels of 'noise' appeared randomly at the second token throughout all the trials.

The experiment was designed with a between-subject factor – Group (Mandarin, English) and within-subject factors – Mode (AO, AV, VO) and SNR (clear, SNR-6dB, SNR-9dB). In each condition, there were 24 tone pairs including 12 pairs of the 'different' type (AB, BA), and 12 pairs of the 'same' type (AA, BB) [3]. The total number of the trail was 216 (3 Modes × 3 SNR × 24 tone pairs). Before the experiment, the participants were given 15 trials for practicing which were not used in the experimental blocks and they were told to respond to the task as quickly and accurately as possible. They were not given any suggestions to pay attention to any specific part of the videos.

2.2. Results

The data results will be reported in discrimination index (DI) ranging from 1 to -1 (see [2]). The overall performance of three modes was subjected to a 3-way mixed ANOVA analysis having a between-subject factor of Group (Mandarin, English) and within subject factors of Mode with 3 levels (AO, AV, VO) and SNR with 3 levels (clear, SNR-6dB, SNR-9dB). Main effect was found on Mode, F (2, 76) = 157.02, $p < .001$ and SNR F (2, 76) = 187.92, $p < .001$. Interaction effect of Mode and SNR was significant, F (4, 152) = 54.58, $p < .001$. Simple effect of Mode showed no significant effect between AO and AV under both clear and noise condition for Mandarin, but for English speakers, their AO (M= 0.32, SD= 0.26) is significantly higher than AV (M=0.21, SD= 0.21) under SNR -9dB ($p < .05$). VO performance was significantly lower than the other modes at clear ($p < .001$) and SNR -6dB ($p < .001$) conditions for both groups. It suggests that Mandarin speakers did not make use of visual cues of AV stimuli to assist tone perception while English speakers relied more on auditory information (visual reduction effect) under higher level of babble noise and the extra visual information lowered their tone performance.

To further explore whether visual effect appears during specific tone pair discrimination, a mixed ANOVA analysis with within factors of 2 Modes (AO, AV), 3 SNR (Clear, -6dB, -9dB) and 6 Tone Pairs (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4) was performed. Main effect was found on SNR, F (2, 76) = 292.14, $p < .001$, Mode, F (1, 38) = 4.89, $p < .05$, and Tone Pair, F (5, 190) = 9.78, $p < .001$. Three-way interaction effect reached significant level, F (10, 380) = 10.48, $p < .0\ 01$. Simple effect analysis on Mode found that Mandarin speakers' AO performance was significantly higher than AV (visual reduction) when they discriminated tone pair T2-T3 ($p < .001$), T2-T4 ($p < .01$) under SNR -6dB and T1-T3 ($p < .05$), T2-T3 ($p < .001$) under SNR -9dB. Visual augmentation was in T3-T4 ($p < .001$) under SNR -9dB. For English speakers, visual reduction effect was found when perceiving T2-T3 ($p < .001$) under SNR -6dB and T1-T3 ($p < .001$), T2-T3 ($p < .01$) under SNR-9dB and their visual augmentation was, again, in T3-T4 ($p < .05$) (see Figure 2).

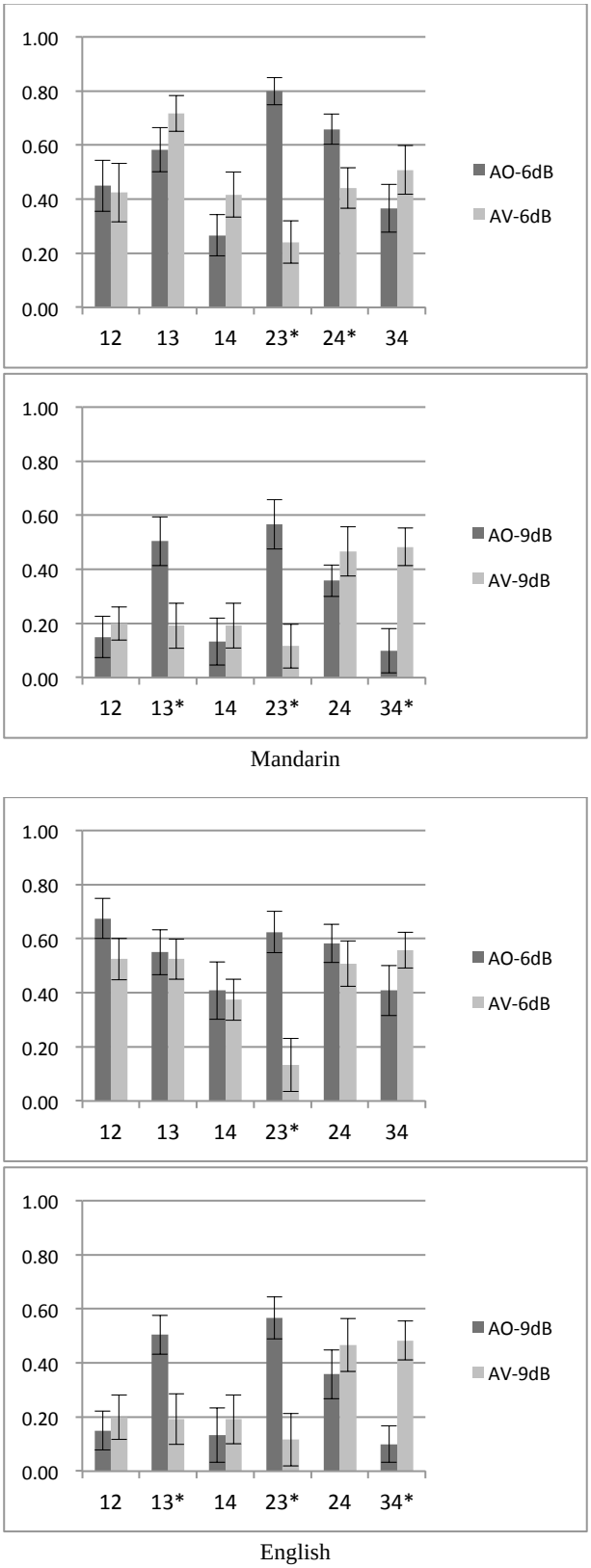


Figure 2. *DI of tone pair in AO and AV mode under SNR -6dB and -9dB for Mandarin and English speakers (*: $p < .05$)*

Although the visual augmentation effect cannot be observed when looking at the overall tone perception, it exists in particular tone pair and was neutralized by the visual reduction effect of some tone pairs. The both groups have a similar pattern of tone pair perception. The speakers benefited from the additional visual information when perceiving the contrast of T3-T4 but not when discriminating T1-T3, T2-T3. The tone pairs having visual reduction effect are more than those having augmentation effect.

Despite the fact that the visual augmentation effect was found in specific tone pair (T3-T4), it was not robust enough to conclude that visual cues can be used to facilitate their tonal perception, therefore we need to test augmentation and reduction effects in the second experiment with a similar AX task.

3. Experiment 2

Experiment 2 employed the same AX task but two tokens in each trial came from 2 different Mandarin native speakers throughout all trials. Unlike experiment 1, two tokens in each trial presented in the same mode (e.g. AO-AO, AV-AV and VO-VO). This adjustment was to avoid participants to use the extra processing load as perceiving stimuli cross channels.

3.1 Methods

3.1.1. Participants

Twenty Mandarin native speakers (age: 22-40; female: 15) and 20 English native speakers (age: 19-20; female: 17) participated in the study. None of them had participated in the first experiment. They had no known hearing problems and had normal or corrected to normal vision. They received payments for their participation.

3.1.2. Materials and Procedures

The same syllable *bai* with 4 tones were presented in 3 modes (AO, AV and VO) and 2 levels of babble noise (SNR-6dB and -9dB) borrowed from Experiment 1 were embedded in the second stimulus in each trial. One speaker was from Experiment 1 material and the additional speaker was a male of Mandarin native speaker. The procedures were identical to Experiment 1.

Table 2. *Acoustic features of stimuli (bai) of speaker 2*

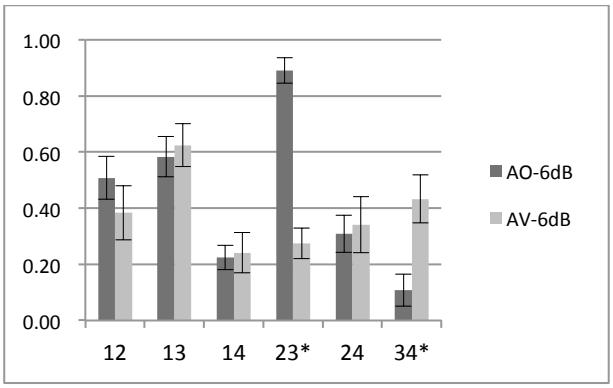
	bai1	bai2	bai3	bai4
Average Pitch (Hz)	135.58	128.75	111.8	140.12
Duration (s)	1.15	0.96	1.12	0.48
Intensity (dB)	64.91	63.89	62.49	65.61

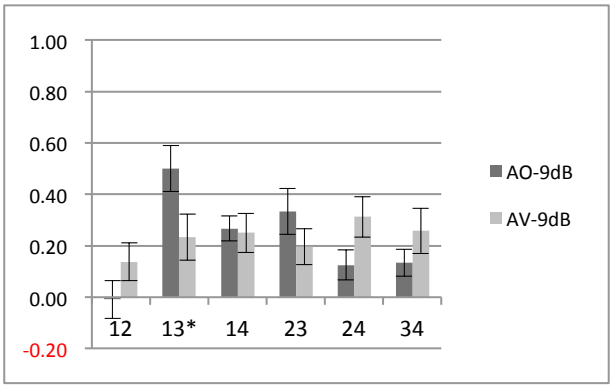
3.2 Results

A 3-way mixed ANOVA with between-subject factor Group (Mandarin, English) and the within subject factors Mode (AO, AV, VO) and SNR (clear, SNR-6dB, SNR -9dB) was performed to test the visual augmentation effect in AV mode. The result finds the significant effect that AO ($M = 0.31$, $SD = 0.13$) is higher than AV ($M = 0.19$, $SD = 0.16$) ($p < .05$) under SNR -9dB among English participants. VO was significantly lower than the other modes at both clear and noise condition

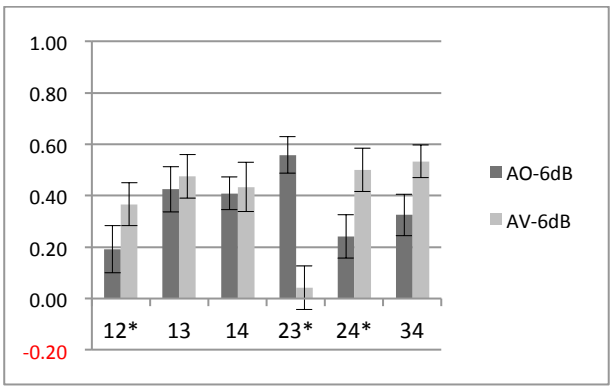
($p < .01$) except for English speakers' performance of AV was as low as VO at the SNR -9dB.

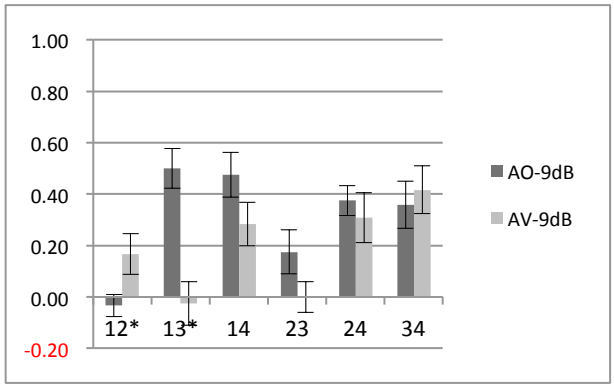
To analyse tone pairs, a mixed ANOVA with a between-subject factor Group (Mandarin, English) and within-subject factors of Modes (AO, AV), SNR (Clear, -6dB, -9dB) and Tone Pairs (T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4) shows the main effect on Group, F (1, 38) = 6.22, $p < .05$, SNR, F (2, 76) = 164.00, $p < .001$ and Tone Pair, F (5, 190) = 9.89, $p < .001$. A 3-way interaction effect was significant, F (10, 380) = 10.63, $p < .001$. Simple effect of Mode finds that Mandarin speakers had the visual reduction effect in T2-T3 (SNR-6dB) ($p < .001$) and T1-T3 (SNR -9dB) ($p < .05$) and visual augmentation in T3-T4 ($p < .01$) at SNR -6dB. The English speakers had the same visual reduction effect as Mandarin counterparts in T2-T3 ($p < .001$) at SNR -6dB and T1-T3 ($p < .001$) at SNR -9dB, and their visual augmentation effect appeared in T1-T2 ($p < .05$), T2-T4 ($p < .05$) at SNR-6dB and T1-T2 ($p < .05$) at SNR-9dB (see Figure 3).





Mandarin





English

Figure 3. *DI of tone pair in AO and AV mode under SNR -6dB and -9dB for Mandarin and English speakers (*: $p < .05$)*

4. General Discussion

From two experiments above, the general visual augmentation was not found under either level of SNR condition for both Mandarin and English speakers, but visual reduction effect as presenting heavy noise (SNR -9dB) during tone perception among English speakers. This is not consistent with our hypothesis and the previous findings. It suggests that English speakers prefer auditory tone to audiovisual tone or that the visual cues mislead their final perception.

However we cannot conclude that visual information would not facilitate their tone audiovisual perception. From the data of tone pair analysis, the extra visual information presented leads to two directions – facilitating or inhibiting lexical tone perception. Based on the data, the both visual effects appeared in different tone pairs. In Experiment 1, the visual augmentation effect was only observed in tone pair T3-T4 discrimination and the visual reduction effect appeared in T2-T3 comparison across all levels of noise for both Mandarin and English speakers (see Figure 2). This indicates that the two visual effects took place in noisy condition, in which the visual augmentation of certain tone pairs was offset by the visual reduction of other tone pairs, thereby neutralizing the final visual effect in general tone perception. In Experiment 2, the pattern of visual augmentation of T3-T4 and visual reduction of T2-T3 can still be observed at the certain level of the noise (SNR -6dB), even though it is not as neat as the results in Experiment 1, for more tone pairs come into play (e.g. visual augmentation in T1-T2, visual reduction in T1-T3) (see Figure 3). This may be due to the task in Experiment 2 is harder than Experiment 1. The tokens in Experiment 2 came from 2 different speakers while in Experiment 1 they were presented from the same speaker. As a result, the participants needed to recognise the visual motion of the tones from different subjects whose individual differences of the pronunciation habits might lead to inaccurate detection of the visual features of the tones.

Regarding the tone pairs that induced visual augmentation effect, the both language groups (especially Mandarin speakers) show that they have greater chance to benefit from the visual cues to discriminate T3-T4 under the noise condition in both experiments, which fits our prediction that the contour-contour tone pairs containing Tone 3 would be

more easily to be discriminated audiovisually, but it is not the case for the contour-contour contrast of T2-T3. The visual effect on this tone pair was reversed (visual reduction effect) in both experiments, which has not been reported by the previous studies. One possible explanation for this may be associated with their acoustic features. For tone 3 and 4, their acoustic duration, pitch height and tone contour are highly different from each other. As can be seen in Table 1 and 2, auditory tone 3 has the longest duration while tone 4 is the shortest one. The pitch height of tone 3 is lowest one whereas tone 4 is the second highest pitch for speaker 1 and the highest pitch for speaker 2. In terms of tone contour, the trajectories of these 2 tones are opposite to each other. This implies that certain acoustic features of tones can be seen from the speakers' faces which were exploited during perceiving tones. Thus the high contrast of these two tones becomes a visually distinctive pair, which facilitates audiovisual perception. When it comes to T2-T3, they have the similar contour shape and duration, albeit the contour-contour tonal contrast. The similarity of their acoustic features leads to the similarity in visual information on the face, which makes T2-T3 to be the most difficult pair to be visually distinguishable, therefore inhibiting perception.

In conclusion, the two experiments provide some evidence that visual lexical tone can be seen and utilised by both Mandarin and English native speakers during audiovisual perception in adverse condition, even though English speakers tend to depend more on auditory cues. Visual tone as another source of information, it can bring an effect that is not necessarily beneficial but inhibited to audiovisual perception, which depends on the similarity between the features of two visual tones. However, our studies have not been able to articulate which visual tone feature is the most critical cue during audiovisual perception. Further research should be focused on addressing this issue.

5. References

[1] Burnham, D., Lau, S., Tam, H., & Schoknecht, C. 2001. Visual discrimination of Cantonese tone by tonal but non-Cantonese speakers, and by non-tonal language speakers. *Proc. AVSP 2001*, Scheelsminde 155-160.

[2] Burnham, D., Ciocca, V., Lauw, C., Lau, S., & Stokes, S. 2000. Perception of visual information for Cantonese tones. In M. Barlow & P. Rose (Eds.) Proceedings of *the Eighth Australian International Conference on Speech Science and Technology* (pp. 86-91), Australian Speech Science and Technology Association, Canberra.

[3] Burnham. D., Attina, V., & Kasisopa, B. 2011. Audio-Visual discrimination and identification of lexical tone within and across tone language. *Proc. AVSP 2011*, Volterra, Italy, 37-42.

[4] Ching, Y. C. T. 1986. 'Voice pitch information for the deaf', in: Engell UCL, et al., ed. Towards better communication, cooperation and coordination, *Proc. First Asia Pacific Regional Conference Deafness*, Hong Kong, 340-343.

[5] Ladefoged, P. 2001. *A Course in Phonetics*, 4th ed. Thompson, Learning, Boston.

[6] McGurk, H., & MacDonald, J. 1976. Hearing lips and seeing voices. *Nature, 264*, 746-748.

[7] Mixdorff, H., Charnvivit, P. and Burnham, D. 2005a. Auditory-Visual Perception of Syllabic Tones in Thai. *Proc AVSP 2005* Vancouver Island, 3 - 8.

[8] Mixdorff, H., Hu, Y., & Burnham, D. 2005. Visual Cues in Mandarin Tone Perception. *Proc. Interspeech. Eurospeech 2005* Lisbon , 405 - 408.

[9] Sumby, W. H., & Pollak, I. 1954. Visual contribution to speech intelligibility in noise. *JASA*, 26, 212-215

[10] Summerfield, A. Q. 1991. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society*. London B. 335, 71-78.

[11] Summerfield, A. Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye*. (pp. 3-51). London: Erlbaum Associates.

1 Thai has 5 tones: mid level (33), low level (11), falling (231), high level (55) and rising (315).

2 A tone contour is pitch movement associated with a particular shape and direction of pitch trajectory. For example, tone 231 can be considered as a contour tone, and contrastively tone 33 is a level tone.

3 Mandarin lexical tones contain 4 tones: Tone 1: high tone (55), Tone 2: rising tone (35), Tone 3: falling-rising tone (214), Tone 4: falling tone (51). Tone 1 is considered as level tone and the others are contour tones. The possible tone pair comparisons are: 1) different type T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4; 2) same type T1-T1, T2-T2, T3-T3, T4-T4.