Auditory-Visual Perception of Mandarin Lexical Tone using 3D Display

Alyssa Dyball

Faculty of Human Sciences

Department of Psychology

Empirical thesis submitted in partial fulfilment of the requirements for the degree of Masters of Research from Macquarie University, on the 9th of October, 2015.

Table of Contents

Page

List of Figures

List of Tables

Abstract

Research on lexical tone perception has identified patterns of head movements that accompany the production of Mandarin tones. The present study used 3D presentations to investigate whether greater access to depth cues could facilitate tone discrimination in native and non-native tone language listeners (Mandarin and Australian-English). Participants completed a same/different task constructed using syllables /ma/ and /na/, spoken with the 4 Mandarin tones by two different talkers. These syllables were presented in 5 modalities: Audio Only (AO), Two-dimensional AV (2D-AV), Three-dimensional AV (3D-AV), Two-dimensional visual only (2D-VO) and Three-dimensional visual only (3D-VO). Within each modality there were two speech production modes: normal speech and clear speech. Contrary to the hypotheses, performance was similar across AO, 2D- and 3D-AV as well as normal and clear speech conditions; neither 3D presentation nor clear speech improved discrimination. However, performance on the visually salient /ma/ was better than for /na/ suggesting that additional segmental visual information facilitated performance. As expected Mandarin listeners performed better than Australian English listeners. The results highlight inter-talker variation in the production of tone and generate new avenues for study including exploring tone perception with hearing impaired listeners who may rely more on and benefit more from 3D presentation.

Declaration of Originality

All works contained within this thesis have not been previously submitted to another university or institution. Ethical clearance for this research was granted by the Human Research Ethics Committee of Macquarie University (reference number: 5201300436).

ADyball.

Alyssa Dyball

9.10.15

Auditory-Visual Perception of Mandarin Lexical Tone using 3D Display

Human communication generally involves face to face interactions and is therefore predominantly multi-modal by nature, composed of auditory, visual and gestural information. This has been recognised in recent technological advances that have focused on providing better ways to approximate the natural communication contexts by providing access to a more comprehensive range of cross-modal information via video chats. While one can still gain a great amount of information from auditory input alone, such as the speaker's mood, size, gender and intent (Ward, 2010), visual information also plays an important role in speech perception.

The interaction between auditory and visual information is elegantly demonstrated by the McGurk Effect (McGurk & McDonald, 1976) whereby a mismatch between the auditory information of one sound and the visual information of another sound results in the percept of a third sound. For example, when listeners *see* a /ga/ being articulated but *hear* a /ba/, most will report hearing a /da/ or a /tha/. This suggests that visual information can alter the auditory perception of speech. Reliance on visual information also increases when listening conditions are not ideal. When listening in noise, being able to see the speaker can increase speech intelligibility, this can be an equivalent of improving the signal to noise ratio (SNR) of up to 22dB, effectively making the speech signal louder and the noise softer (Sumby & Pollack, 1954). Similarly, when the speaker is visible, listeners are able to perform better in speech detection tasks with higher levels of masking noise (Grant & Seitz, 2000).

However, speech is composed of both segmental information (e.g., vowel and consonants) and suprasegmental information (e.g., sentence intonation, stress, and focus). Since early work on the McGurk Effect, many studies have examined the role of auditory-visual (AV) information on these different levels of speech.

1.1 The Role of AV Information in the Perception of Segments

There is now a vast body of research investigating the visual contributions to segmental speech perception. One such study, by Wang, Behne and Jiang (2009), investigated the perception of place of articulation (POA) for English fricatives, e.g. labio-dental: /fi/, interdental: /thi/ and alveolar /si/ in auditory only (AO), auditory-visual (AV) and visual only (VO) conditions. Overall, performance on a discrimination task was better in AV followed by AO and then VO conditions. However, discrimination performance was only better in AV conditions for labiodentals and inter-dental fricatives which are visually distinctive but not for alveolar fricatives which are not visually distinctive. This suggests that the discrimination of POA is facilitated by the addition of visual information.

Similarly, studies on vowel perception found an interaction between auditory and visual information in speech perception. An early study by Summerfield and McGrath (1984) reported that participants were able to identify incongruent pairings of auditory and visual vowel recordings. Furthermore, an identification task in the same study revealed that as the visual salience of the vowel decreased, e.g., /oo/ greater visual information and /ee/ less visual information, it was more likely to be identified according to auditory information and less likely to be influenced by visual information. Together these studies suggest that visual information plays an important role in the perception of both consonants and vowels. However, the amount of visual information inherent in the production of these segments influences the final integration of auditory and visual speech.

In a cross-linguistic context, the addition of visual information can also be helpful for listeners when distinguishing between non-native speech contrasts. Studies investigating non-native perception of consonant-vowel (CV) syllables in Japanese and English revealed that listeners who are unfamiliar with the language relied more on visual information than native listeners for speech discrimination, which improves their perception of non-native phonemes (Sekiyama & Tohkura, 1993). Similar improvements in AV conditions were reported in Spanish-speaking second language (L2) learners of English (Ortega-Llebaria, Faulkner & Hazan, 2001) and Mandarin and Thai tone language listeners (Chen & Hazan, 2007). A subsequent study found that in addition to greater reliance on visual information, the facilitative effects of visual information for non-native listeners may vary across different phonemes, with some visual information being detrimental to intelligibility (Kawase, Hannah & Wang, 2014). Nevertheless, such results demonstrate that L2 learners are also capable of using visual information to discriminate non-native phonemic contrasts.

Other research using electrophysiology (EEG) also supports the important role of visual information in speech perception. An EEG study by van Wassenhove, Grant and Poeppel (2005) found that visual information facilitates faster cortical processing of CV syllables. Such findings argue for an early integration of the auditory and visual signal. However, in line with the above behavioural studies, the temporal facilitation of the cortical processing was dependent upon the visual salience of the syllable. For example, in the visual only condition the visually salient syllable /pa/ was correctly identified 95 % of the time and with a shorter latency response than the less visually salient /ka/, which was correctly identified at only 65% of the time. Therefore, visual salience is important for both accuracy and speed of speech processing.

While these findings suggest that visual information plays an important role in segmental speech perception, other researchers have explored the use of visual information in suprasegmentals, especially those related to changes in pitch.

1.2 The Role of AV Information in Pitch Perception

In English, pitch changes are associated with a range of suprasegmental functions. These include sentence intonation (e.g., questions and statements), contrastive focus (e.g., RED ball vs. red HAT), and lexical stress (e.g., conVICT vs. CONvict, as verb vs. noun). A study by Cvejic, Kim and Davis (2010) examined two of these prosodic contrasts, focus and sentence intonation, in English. Listeners were able to match silent video recordings of sentences which shared the same prosody, e.g., two different tokens with question intonation, as well as match audio recordings of sentences to the correct video counterpart with good accuracy. As the videos were silent, this suggests that head movements provide reliable visual information that conveys the modulation of pitch characteristic of sentence level prosodic features. A production study in French also reported an association between eyebrow raising and head movement for focus (Dohen, Lœvenbruck & Harold, 2006) with some variation across talkers. Nevertheless, these studies confirm the importance of head and facial movements as reliable cues for conveying pitch information. However, this raises questions about the role of visual information for other uses of pitch, such as lexical tone.

1.3 Lexical tone

When pitch is used to distinguish word meaning, it is referred to as lexical tone. Lexical tone is a feature of many languages, such as those spoken in China, presenting an additional layer of complexity beyond vowel and consonant contrasts. Acoustically, lexical tone is defined by changes in the fundamental frequency (f0) over the syllable, which is perceived as pitch. Here, the perceived pitch height and contour, as well as duration, effect the perception of lexical tone (Massaro, Cohen & Tseng, 1985). Thus, each specific tone is identified by its distinct f0 pattern produced on the sonorant part of the syllable (i.e. vowels and nasal consonants) (Gandour & Harshman, 1978). Each tone language has a unique tone system. One such tone language is Mandarin, the official language of mainland China, which has 4 lexical tones: flat, rising, dipping, and falling. These tones are used by talkers to alter word meaning. For example, the Mandarin word /ma/ can be produced on four tones as four different lexical items, i.e., flat tone meaning *mother*, rising tone meaning *hemp*, dipping tone meaning *horse*, and falling tone meaning *to curse.*

Adult native language listeners of tone languages perform at near ceiling levels on tone identification tasks in normal listening conditions (Mixdorff, Hu & Burnham, 2005), predicted by their daily use of relevant pitch information in the production and comprehension of speech. However, in adverse conditions, listeners often encounter difficulties~~.~~ A study investigating the effects of different noise levels on vowel intelligibility and tone recognition found that while performance in conditions masked by white noise remained above 60%, when speech was masked by multi-talker babble at a level of +25 dBA, simulating a noisy crowd situation, Mandarin listeners were unable to identify tone at a level above chance (Dees, Bradlow, Dhar & Wang, 2007). While tone recognition was impaired, participants' ability to identify vowels remained intact. Another study found similar results (Wang, Jongman & Sereno, 2001, suggesting that lexical tone is not perceived like consonant and vowel segments despite it occurring over segments. These results from auditory only studies raise questions about whether visual information is important for lexical tone processing.

1.3.1 AV Information for Lexical Tone. Some recent research investigated the use of visual information in lexical tone perception. An initial study by Burnham, Ciocca and Stokes (2001) investigated the auditory-visual processing of Cantonese tones, a tonal language widely spoken in Southern-China. Participants were required to listen and/or watch spoken Cantonese stimuli and identify the correct word and tone on a screen. Native

Cantonese listeners completed this task in three conditions: AO, AV and VO. Participants were either trained phoneticians or untrained listeners. Overall, performance in both the AO and AV conditions were similar. Phonetically trained participants performed better than untrained listeners, but only when the auditory information was provided. However, untrained listeners were able to discriminate between tones at a level above chance when only the visual information was provided. This may be explained by the relative importance given to auditory information by trained phoneticians. These results provide some evidence for the existence of visual information for lexical tone.

In a follow-up study (Burnham, Lau, Tam & Schoknecht, 2001), native Thai- and Australian-English (AusE) listeners were tested using the same stimuli. A babble-masked condition was also included to explore the use of visual information in noisy contexts. In each trial participants listened and/or watched 2 stimuli and were asked to decide whether the tones were the same or different. Consistent with the hypothesis that experience with tone would predict better performance, Thai listeners performed better overall on tone discrimination than AusE listeners. In clear audio, performance was the same across AO and AV conditions for both groups. However, as the auditory information became less reliable with the addition of babble noise, the addition of visual information improved performance.

Using a slightly different paradigm, Mixdorff, Hu and Burnham (2005) investigated AV perception of Mandarin tones. Participants were required to identify the tone they had heard in both clear and noise-masked AO and AV conditions. To direct the listeners’ focus to the articulatory movements of the lips and chin, only the lower half of the talker’s face was presented. In the clear auditory conditions, visual information did not improve performance, but it did improve performance in the babble-masked condition and even more so as the SNR decreased. Similar advantages were found when AV stimuli were presented with auditory noise in Thai (Mixdorff, Charnvivit & Burnham, 2005) and in Vietnamese (Mixdorff, Lương, Nguyen & Burnham, 2005). However, most of the listeners in the Mixdorff, Charnvivit & Burnham (2005) and Mixdorff, Lương, Nguyen and Burnham (2005) studies were familiar with the talker; thus familiarity with the talker could have contributed to better performance in the conditions with visual information.

Expanding on these findings, a study by Chen and Massaro (2008) further explored visual tone in Mandarin. More specifically, their study aimed to determine whether visual information alone is informative and useful for the identification of Mandarin lexical tones. This study took place over 3 days. On the first day, participants were recorded as they produced Mandarin tones. From the recordings of all of the participants, informal descriptions of the specific movements which accompanied each of the lexical tones were created. On the second day, participants were required to watch visual only recordings of the stimuli and identify the word and tone produced. The stimuli in this task included both stimuli derived from their own productions and from the other participants. On the final day, participants received training using the descriptions of the visual movements observed by the researchers. Following the training session, participants performed the same identification task as they had on the previous day. Comparison between performance on pre and post training revealed that performance improved on the visual only identification of tones from a level close to chance to significantly above chance. Auditory training over a longer period, spread over two weeks, has also been effective with non-native listeners for perceiving Mandarin tones (Wang, Spence, Jongman & Sereno, 1999). The results from these two studies raise the possibility that directing attention to visual information may facilitate faster learning and enhanced performance on tone discrimination, over and above auditory only input. However, it is possible that in the Chen and Massaro (2008) study researchers may have inadvertently induced exaggerated speech articulations by instructing the talkers to “speak clearly and distinctively” (pp. 2358). Thus it is also possible that exaggerated speech may be accompanied by enhanced visual information that facilitated the identification of lexical tone.

In a parallel set of production experiments, visual tone information was studied using sensors which tracked head and facial movements during speech production. These studies used the OPTOTRAK technology (OPTOTRAK, Northern Digital Incorporated), a system that tracks kinetic motion through small sensors placed on the head and face. The data from these sensors are then plotted in a 3D space, allowing researchers to track rigid (global head movements) and non-rigid (facial muscle movements) head movements across time. Finally, principal components analysis (PCA) is conducted on the data to identify movements that may characterise the production of a particular tone. Finally, by correlating these movements with f0 and determining how they change across time, sets of visual movements that relate to the acoustic production of tone can be determined.

The first of these two production studies investigated the visual information specific to Cantonese, which has a 6 tone system (Burnham, et al., 2006). A female native-speaker was recorded producing Cantonese words for each tone in isolation and in sentences. The PCA revealed that there are approximately 14 key movements, including both rigid and non-rigid head movements. Correlation between f0 and these movements across time revealed that rigid rather than non-rigid movements contribute most uniquely to lexical tone production. A subsequent perception task using the data extracted from the OPTOTRAK output found that rigid movements were most useful for the perception of lexical tone while non-rigid movements were most useful for the perception of vowels and consonants.

A second production study investigated visual tone for Mandarin (Attina, Gibert, Vatikiotis-Bateson & Burnham, 2010). This study also recorded a female native-speaker as she produced Mandarin words carrying the four tones in isolation and in sentences. In comparison to Cantonese, this study identified a smaller number of key movements, 6 rigid and 6 non-rigid, in Mandarin tone production. This suggests that visual movements are specific to individual tone language systems and the number of movements may be associated with the complexity of the system. However it must be noted that this study used guided PCAs and this may have contributed to the finding of fewer key movements.

Unique patterns of correlations were found for each tone between non-rigid head movements and f0. For example, the dipping tone correlated most strongly with a back and forth head movement, which the authors suggest may be involved in the modulation of pitch contour. Similarly, the flat tone was correlated with a nodding movement which may be related to the production of a high tone. These results suggest that over and above the existence of visual information that accompany lexical tone, each tone may have its own unique set of visual movements that help distinguish it from other tones.

Building upon these findings, a recent study by Smith and Burnham (2012) investigated how visual information could be used in applied settings. Using a Matlab toolbox, audio stimuli were generated that closely simulated what a cochlear implant (CI) would deliver to a recipient. The lexical tone stimuli created using the cochlear implant simulator in AO and AV conditions were compared to stimuli without any manipulations in AO, AV and VO conditions. Both Mandarin- and Australian English (AusE) listeners were tested. As expected, discrimination was poorer using the CI simulated stimuli than with stimuli without manipulations. However, the addition of visual information in both types of stimuli improved performance. Discrimination was also above chance level when participants were provided with only visual information. Interestingly, AusE listeners outperformed Mandarin listeners in VO discrimination; suggesting that native listeners may be paying less attention to visual information for lexical tone.

Together, this growing body of research provides evidence to support the existence of visual information related to lexical tone production and that visual information is of greatest use to listeners in situations where the auditory information is degraded by noise. However, in reality, human listeners inadvertently employ a number of strategies to alter their manner of speaking, in order to compensate for the loss of quality in adverse listening conditions.

1.4 Speech Adaptation

In adverse listening conditions, talkers modify their speech in an attempt to increase intelligibility. This is typically found when speaking to an individual with a hearing impairment (HI). This form of adaptation, referred to as clear speech, is characterised by a slowed speaking rate and increases in intensity, word duration, and number of pauses (Picheny, Durlach & Braida, 1986). Clear speech has been demonstrated to help to increase speech intelligibility for listeners with HIs as well as people with normal hearing in both AO (Payton, Uchanski & Braida, 1994) and AV conditions (Helfer, 1997). A study that induced clear speech, in which talkers were required to communicate with a listener who was listening to babble noise, reported visually exaggerated vowel articulations in English (Hazan & Kim, 2013). This exaggerated visual information might facilitate speech intelligibility in clear speech. However no one has yet examined this specifically in relation to lexical tone perception.

Given the reported importance of rigid head movements in the production of lexical tone, traditional 2D presentations of speech will be a serious limitation in presenting the full range of visual movements, especially those with a forward and backward movement which rely on depth cues. This is also one potential reason for why visual information is under used by native listeners from previous studies. In order to increase access to these cues, the present study will make use of new 3D video technology.

1.5 3D Perception

In face to face interactions, people view the talker's face from multiple viewing angles afforded by 3D visual perception. The most obvious difference between real-world speech interactions and the representation seen in the lab using 2D images is depth cues. The added richness in depth cues provided by 3D presentation may provide access to the necessary visual information that accompanies the production of lexical tone. It is expected that this will be particularly useful for the perception of the rigid head movements previously identified, for example the forward and backward movement for the dipping tone in Mandarin.

This technology is now widely available as a commercial product and has found a great many uses in a wide variety of contexts including medicine (Wen, Markey & Muralidlhar, 2014). This study aims to use 3D displays to investigate the usefulness of visual information for the discrimination of lexical tone. If successful, this research could have implications for teaching second languages and even clinical applications such as intervention programs for people with hearing impairment, by improving the amount of visual information accessible by the listener.

1.6 The Present Study

In the present study we aim to investigate whether 3D presentation can help to improve lexical tone discrimination by increasing access to depth cues. As a secondary aim, we will investigate whether clear speech, simulating speech to impaired hearing listeners (produced while listening to babble noise) will enhance lexical tone perception.

1.6.1 Hypotheses. First, informed by the results of previous studies, we expect the discrimination performance for lexical tones to be best in the AV conditions, followed by the AO conditions and worst on the VO conditions.

Second, since viewing in 3D is the default mode of perception, it is more realistic and provides a more comprehensive range of visual information so listeners are not required to rebuild the scene using 2D information, we predict that discrimination performance on lexical tone will be better using 3D compared to 2D stimuli.

Third, we expect that tones elicited during clear speech, with enhanced visual movements, will facilitate tone discrimination compared to normal speech and performance will be better in the clear speech condition compared to normal speech.

Fourth, overall performance for Mandarin (native tone language) listeners should be better than AusE (non-native tone language) listeners across all stimuli presentation conditions, due to native language experience with lexical tone.

Finally, as an exploratory hypothesis, we will also investigate whether the visual salience of the consonant (e.g. visually salient /ma/ vs. visually ambiguous /na/) might influence the perception of lexical tone. Since the /ma/ syllable contains more visual information, both for the consonant and tone, it is expected that this will facilitate visual speech perception. Thus, overall performance on /ma/ should be better than /na/ in AV and VO conditions.

2. Method

2.1 Participants

A total of 22 native listeners of Australian-English, mean age 19.85 years (range = 18-30; SD = 3.10) were recruited from a pool of psychology students and received course credit for their participation.

A total of 22 native listeners of Mandarin-Chinese, mean age 24.21 years (range = 19-37; SD = 4.78) were recruited externally via advertisements placed around the university campus, advertisements on Facebook and via email. These participants were given a small cash payment ($15 AUD) as reimbursement for their time. All Mandarin listeners completed schooling in Mainland China and reported continued daily use of Mandarin.

All participants had self-reported normal hearing and language abilities. Professional musicians, defined as having begun playing a musical instrument prior to age 7, with at least 10 years of experience and consistent practice at least 3 times a week (Parbery-Clark, Skoe, Lam & Kraus, 2009) were excluded from participation due to the association found between musical ability and enhanced pitch perception abilities (Kraus & Chandrasekaran, 2010).

Prior to the onset of the experiment, participants were screened using the Stereo Fly Test (Stereo Optical Co., Inc., Illinois) a test of stereoscopic depth perception. This was included to control for natural variation in stereoscopic vision, which studies have attributed to differences in visual acuity between the two eyes (Lam, Chau, Lam, Leung & Man, 1996). Results on this test indicated good stereoscopic acuity (≤ 60 arcsec) (Brooks & Rafat, 2015) for most participants with the exception of 3 English and 3 Mandarin listeners whose results were excluded from subsequent analysis.

All participants received written instructions in their native language, i.e., English for English listeners and Simplified Mandarin for Mandarin listeners. Written consent was obtained prior to the commencement of the experiment. Ethics approval was granted by the Human Research Ethics Committee (Medical Sciences) of Macquarie University, Sydney.

2.2 Apparatus

Video recordings were captured using a Fuji Finepix Real 3D camera. This camera uses a dual lens system that simultaneously captures both right eye and left eye streams in 1280 by 720p at 24 frames per second. The distance between the two lenses is 75mm which approximates the average human inter-ocular separation (distance between left and right pupils) 65mm (Snowden, Thompson & Troscianko, 2006). The talker was seated .5m from the white backdrop and the camera was mounted on a tripod that stood at eyelevel, at the recommended filming distance of 1.3m from the talker.

As the video camera could only provide a compressed low quality recording, the audio was simultaneously captured by an external directional microphone that was held by a tripod placed at a distance of .5m from the face. The external audio was digitised at 44.1 kHz with a sampling rate of 16 bits high quality uncompressed recording. During recordings a clapper board was used to allow for the synchronisation of the high quality audio recording and that captured by the video camera during subsequent editing.

Both 3D and 2D stimuli were displayed using a True3Di SDM-240 24″ stereoscopic 3-D monitor. This monitor used 2 linearly polarised LCD screens positioned perpendicular to one another with a mirror positioned between at an angle of 45^{o} to create a 3D image. Both screens had a resolution of 1920 × 1200 pixels and were frame synchronised at 60 Hz. The experiment was run using a PC running Windows XP 32-bit with an Intel Core i7-980X 3.33 GHz processor, 8 GB RAM, and two NVIDIA Quadro FX 5000 graphics cards.

During the experiment, a chin-rest was used to control both the viewing position, level with the middle of the screen and the viewing distance, set at 130 cm from the screen surfaces to match the original filming distance. The final video was also cropped to match the real-world size of the talkers’ faces. These control measures aimed to approach orthostereoscopy, or to attain the closest match possible between the retinal images created by our 3D stimuli and those of the original real world 3D stimuli that we captured during filming.

In 3D viewing conditions participants wore linearly polarised 3D glasses that work by simultaneously providing both left and right streams to the viewer (see Figure 1).

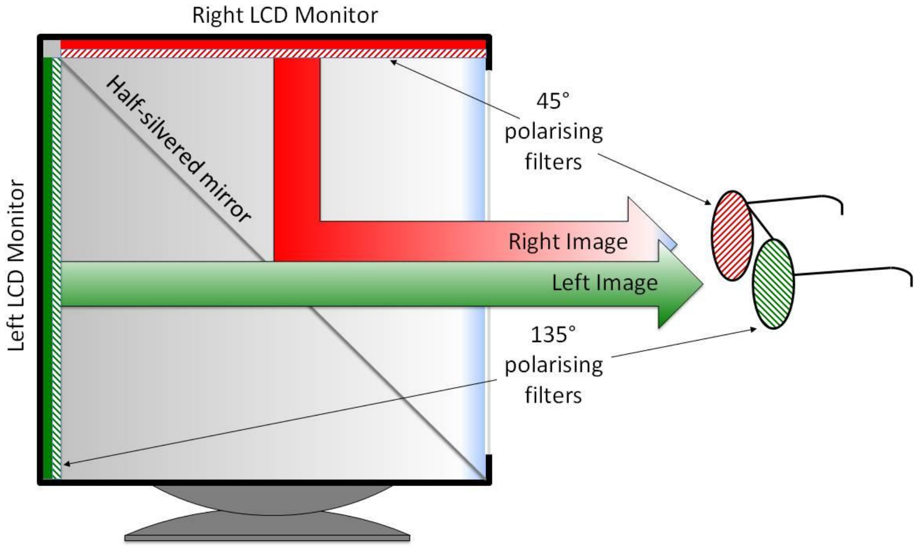

Figure 1. Set up for the 3D conditions. This figure demonstrates the configuration of the True3Di monitor and the 3D glasses with different lens for left/right eyes.

In order to control for the effect of wearing glasses, a pair of glasses which converted the 3D image back to 2D was created for use in 2D viewing conditions. These 2D glasses were constructed by flipping the right eye lens from a pair of 3D glasses around the vertical axis. This removed all 3D effect by restricting the visual input to that displayed on the left monitor (see Figure 2).

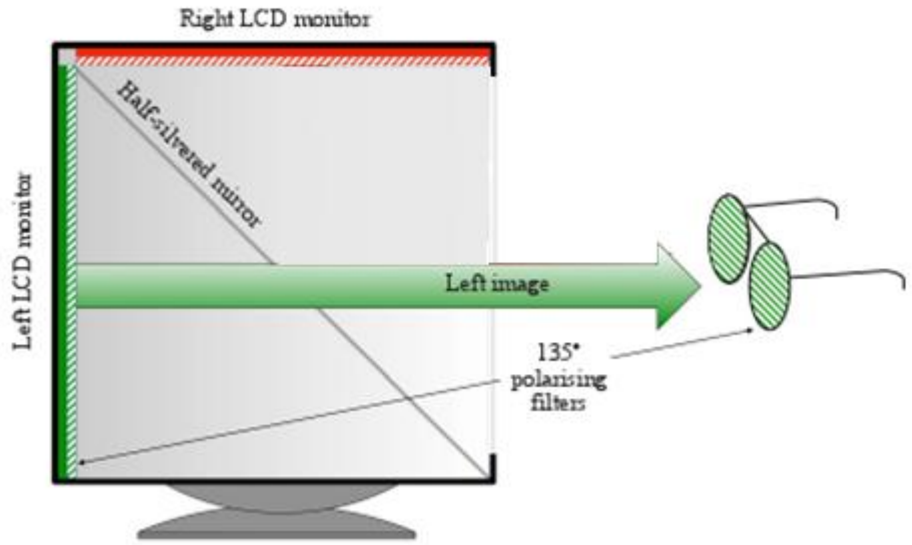

Figure 2. Set up for the 2D conditions. This figure demonstrates how the 2D glasses, using 2 rotated lenses, only provide visual input from the left LCD monitor.

2.3 Stimuli

Two female native talkers of Mandarin Chinese from Beijing who completed university level schooling in Mainland China, aged 24 and 31, were recruited from lab staff. Both reported continued daily usage of Mandarin.

The syllables produced by the talkers were /ma/ and /na/ on all four Mandarin tones. These syllables were chosen to allow for a place of articulation contrast, bilabial /m/ versus alveolar /n/, and also to allow comparison between a visually distinctive bilabial /m/ versus the less visually distinctive alveolar /n/. Furthermore, the syllables /ma/ and /na/ were chosen as all four tones carried on these syllables represent meaningful words in Mandarin with the exception of /na/ on tone 1. See Table (1) for a summary of the stimuli and the English translation.

Table 1

Summary of stimuli and English translations

Syllable	Tone	Meaning
/ma/	1	mother
	2	hemp
	3	horse
	4	scold
/na/		
	1	-
	2	take
	3	what
	4	that

2.3.1 Talker Recording. Recordings were carried out in a sound-attenuated video recording booth lit with a flood light (tri-colour light). A white backdrop was used to create contrast with the talkers who wore black t-shirts. Their hair was tied back and off the face. Talkers were seated during the recordings and were instructed to look directly into the video recorder lens while producing the syllables and to ensure that their expression remained neutral.

In order to mimic a conversational interaction, the talkers were required to speak directly to a listener, who sat behind the camera with her face visible to the talker. The listener was a female native-speaker of Mandarin. She reported continued daily usage of Mandarin.

To ensure that the talker was speaking intelligibly, the listener was asked to identify the syllable and tone spoken by the talker. This was done by the listener showing the syllable heard from a list written in Pinying displayed on a tablet. Pinying is taught in China using as system of Latin alphabet with tone markings to represent spoken syllables. If the listener identified the correct syllable and tone, the talker proceeded to the next syllable.

Recordings were completed in two conditions, quiet trials and babble trials. During quiet trials the listener listened to the talker in silence. During the babble trials the listener wore Sennheiser High Definition enclosed headphones and listened to multi-talker Mandarin babble, such as that previously used by Lee, Tao and Bond (2010). The babble was constructed from recordings of 5 different Mandarin speakers and was played at a sustained level of 30 dBA. During the quiet trials, the talker was instructed to speak so that the listener could understand, as they would during a normal conversation. During the babble trials the talker was informed that the listener was listening to loud babble noise and again instructed to speak so the listener could understand. This was intended to induce ‘clear speech’ (Picheny, et al., 1985) from the talker and to simulate an interaction between a normal hearing person (with self-feedback intact) and a hearing impaired person. Across both conditions, only one production from the talker was necessary for the listener to make a correct identification of the syllables; no syllables required a repeat.

An original practice session of 30 minutes was carried out to familiarise the talkers with the recording protocols. In a following session, each syllable was produced 3 times in random order. Quiet trials were always recorded prior to the babble trials. Trials with visual or auditory artefacts such as smiles, large irregular eye-blinks or auditory noise were discarded.

Measurements of the duration and f0, at onset, mid and offset, of the recordings were taken using speech editing software (Praat, version 5.4.08). These measurements were compared across the normal and clear speech recordings using two individual repeated measures ANOVAs; the tokens for each talker were analysed separately. Overall, the clear-speech recordings were significantly longer in duration, $F(1, 46) = 171.971$, $p < .05$ and significantly higher in pitch, $F(1, 46) = 8.982$, $p < .05$. Thus, the manipulation used in the present study was sufficient to induce changes in the production of tone and syllable duration. See Table (2) for a summary of the durations and Figures (3 and 4) for the f0 values of the final stimuli, separated by syllable e.g. /ma/ and /na/. Figures (3 and 4) were constructed by extracting the f0 at 10 time points from the onset to offset of each syllable and plotted over time.

Table 2

Durations of final stimuli /ma/ and /na/

Syllable	Tone	Talker 1		Talker 2	
		Normal	Clear	Normal	Clear
/ma/	1	.633	.644	.412	.587
	2	.672	.858	.492	.592
	3	.693	.817	.536	.597
	4	.491	.629	.356	.379
/na/	1	.625	.735	.469	.556
	2	.668	.757	.481	.558
	3	.660	.987	.516	.661
	4	.602	.613	.298	.339

Note. All measurements are in seconds.

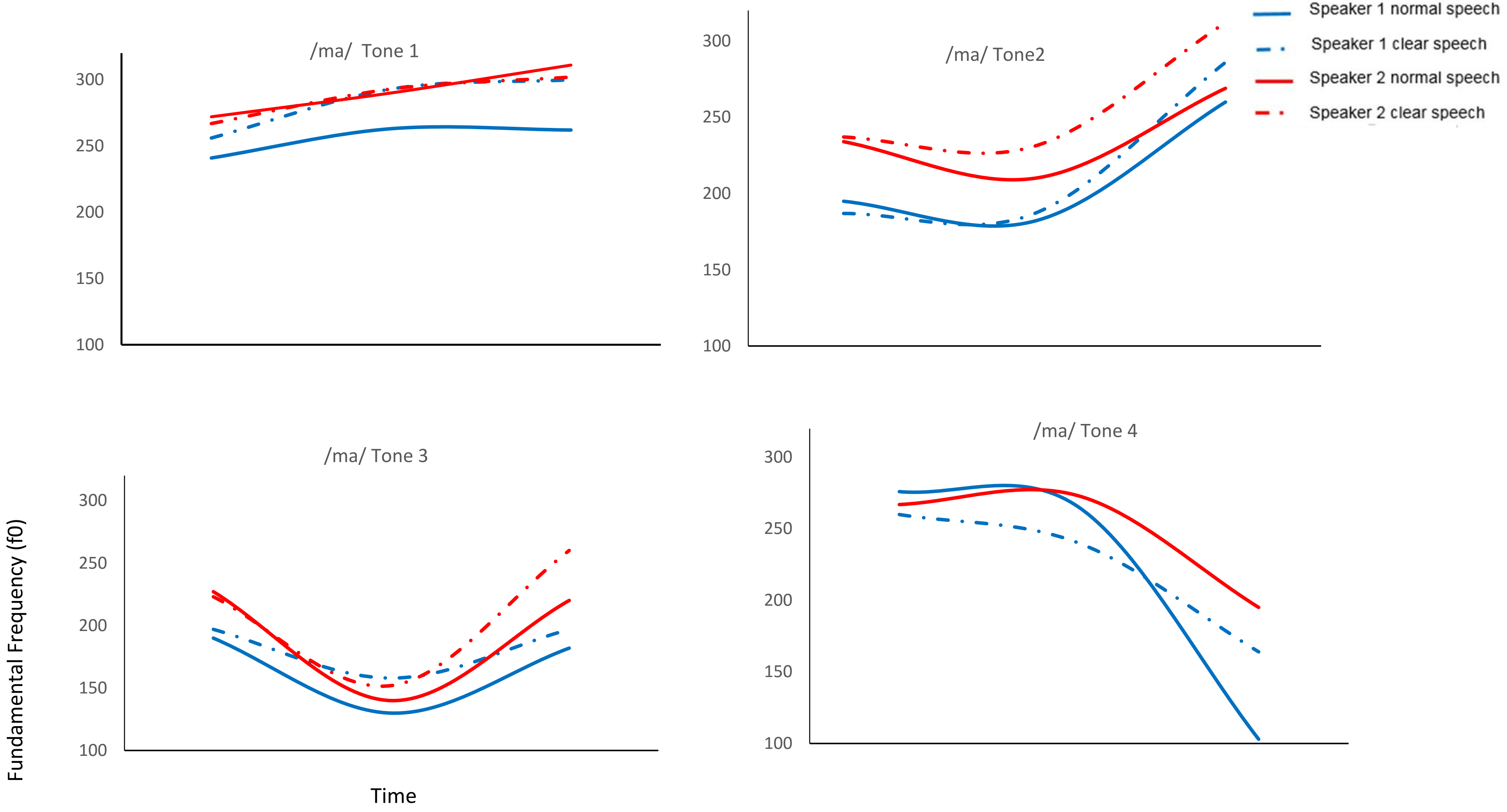

Figure 3 Pitch plots of /ma/ spoken with all four tones. The horizontal and vertical axes represent time and f0 respectively.

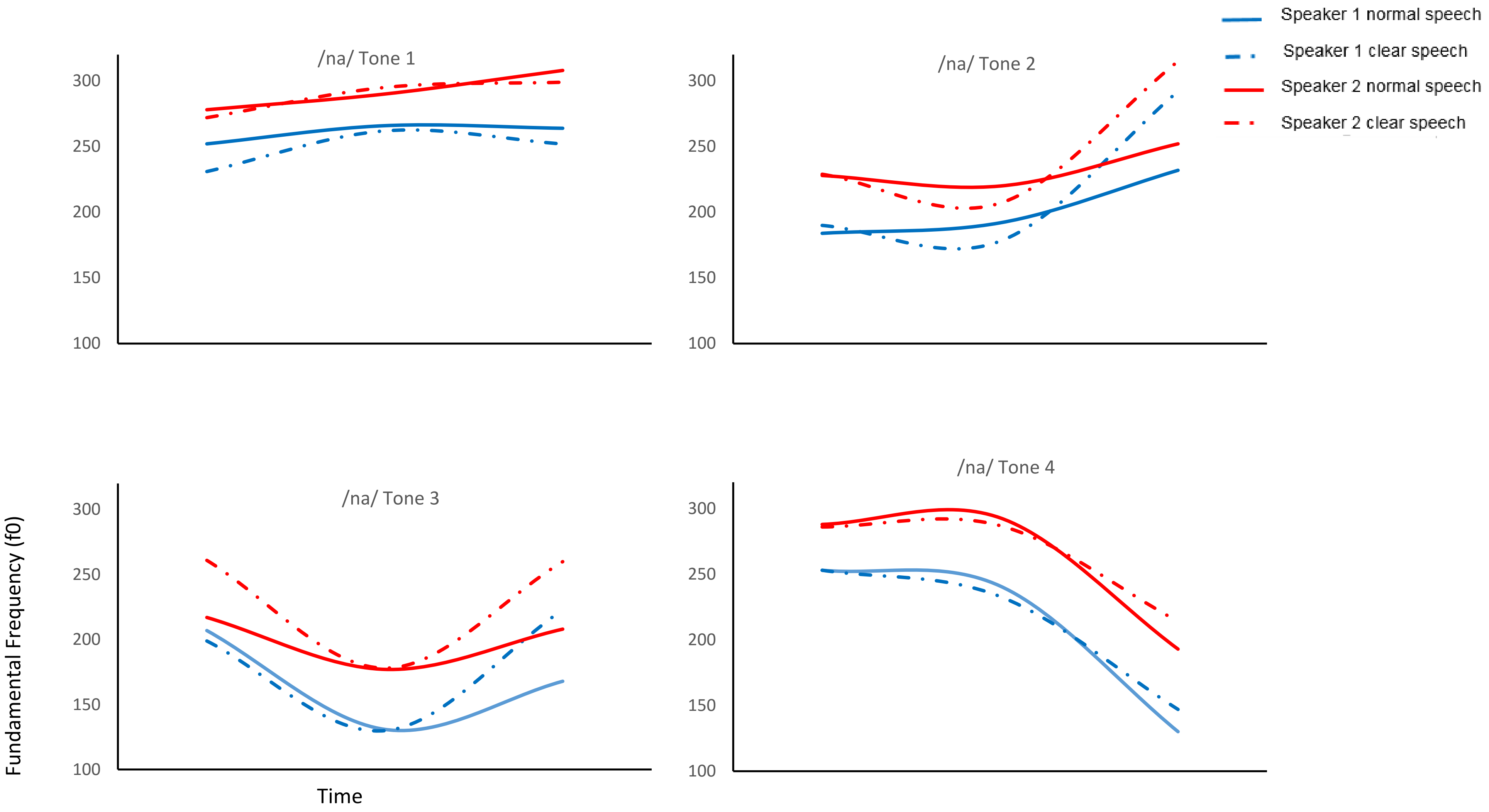

Figure 4 Pitch plots of /na/ produced with all four tones. The horizontal and vertical axes represent time and f0 respectively.

2.3.2 Video Editing. The speech recordings were segmented offline using Praat (version 5.4.08). Onset of the syllable was marked at the onset of the nasal and offset was marked at the cessation of the waveform. The syllables were then extracted from the long waveforms and scaled to 70 dB to equalise intensity across talkers. This level is a comfortable listening level commonly employed in speech perception experiments (Takata & Nábělek, 1990; Wang, Behne & Jiang, 2008). Two hundred milliseconds of silence was then added to either side of the target syllable and the final auditory clips were saved as uncompressed wave files.

The 3D footage was processed offline using video editing software (Sony Vegas Pro, version 10). Firstly, the external audio was synced with the audio stream from the video by identifying and aligning the auditory waveforms using the clapper board as reference points. The audio from the video recording was subsequently removed from the videos for stimulus presentation.

Using the Sony Vegas autocorrect function, small differences between the left and right streams in vertical alignment, rotation, zoom and key-stoning were corrected, to prevent errors during the perception task caused by inaccurate representation of depth cues (Santoro, AlRegib & Altunbasak, 2012).The clips were then horizontally realigned using the nostrils as the reference point. This was done to ensure that the face appeared to naturally extend forwards (in 3D) from the neck at the bottom of the screen and to prevent cue conflicts, caused by clash between binocular disparity cues, which indicate that the neck is in front of the screen, and monocular occlusion cues that indicate that the neck is behind the screen (Sato & Howard, 2001). Such violations can produce confusing visual percepts, e.g., a floating head. The clips were then cropped to closely approximate the actual head size of the talkers when presented onscreen. The final 3D videos were saved in side by side format (with left and right streams joined side by side), in an avi video container.

2.3.3 Rating of Stimuli. Two additional Mandarin listeners from Beijing rated the auditory recordings. They were asked to identify both the syllable and tone, and give the production a rating between 1-6 on a scale of naturalness, 1 being unnatural and 6 being very natural. One hundred percent of the syllables were correctly identified by raters and were given either a 5 or a 6 rating of naturalness. From these stimuli the production closest to the average duration for the talker and tone were selected for the final AV stimuli. This was to ensure that talker rate and mean duration of tones were maintained to reduce within talker and tone variations in the data.

A further 2 Mandarin listeners from Beijing rated the finished AV stimuli. They too were asked to identify the syllable and tone and give a rating of naturalness. To ensure that the audio and video were not out of synchronisation, these raters were asked to pay particular attention to the match between mouth movement and speech sound. All of the final stimuli were successfully identified and rated natural (received a rating of either 5 or 6). A measure of inter-rater reliability, using a 2 way random intra-class correlation, was performed to determine the consistency of the ratings. The agreement between the ratings was high (Shrout & Fleiss, 1979), for single measures ICC= .72 ($p < .001$) 95% CI (.50, .85) and averaged across measures ICC = .84 ($p <. 001$) 95% CI (.67, .92).

2.4 Design

The experiment used a mixed design with one between subjects factor of language on two levels: Mandarin and Australian-English (AusE). There were several within subjects independent variables (IV): Input mode (5 levels: AO, 2D-AV, 3D-AV, 2D-VO, 3D-VO), tone contrast (10 levels: (/Tone1/+/Tone1/, /Tone1/+/Tone2/, /Tone1/+/Tone3/, /Tone1/+/Tone4/, /Tone2/+/Tone2/, /Tone2/+/Tone3/, /Tone2/+/Tone4/, /Tone3/+/Tone3/, /Tone3/+/Tone4/, /Tone4/+/Tone4/) to help identify which contrasts were most confusable for the two groups of participants, speech mode (2 levels: normal speech and clear speech) and syllable (2 levels: /ma/ or /na/).

In summary, each tone contrast was presented twice in each input mode by each speech mode. Thus 5 input modes x 2 speech modes x 10 contrast pairs x 2 repetitions resulting in a total of 200 trials per session.

One dependent variable was measured, percentage correct, which was included as a measure of accuracy.

2.5 Procedure

Stimulus presentation was controlled by E-prime (version 2.0). Audio was delivered using Sennheiser High Definition enclosed headphones. The intensity of the audio was calibrated using a sound-level meter to an average level of 65 dBA. The intensity level was checked prior to the beginning of each testing session to ensure levels had not been adjusted.

Participants were asked to complete an AX or a same/different task. Two stimuli were presented, and at the end of the second stimulus, the participants were asked to indicate via a button press whether the stimuli were the same or different. The stimuli were separated by an interstimulus interval (ISI) of 1000ms. The choice of ISI was informed by research that found that AX tasks with shorter ISIs (500ms or less) encourage reliance on short-term memory stores of the stimuli and resulted in discrimination based on acoustic level processing (Werker & Tees, 1984). Longer ISIs in comparison encouraged more phonemic based discrimination as short-term memory stores of the acoustic features of the stimuli had already decayed (Werker & Tees, 1984). Additionally, the buttons for same and different were switched for half of the participants, in order to control for response bias.

The experiment had both same and different trials. In order to control for participants using idiosyncratic movements of the talkers to perform the task, during each trial participants heard 2 syllables produced by 2 different talkers. During same trials participants heard the same syllable, e.g. talker 1 /ma1/ vs. talker 2 /ma1/. However, on different trials the syllable varied in tone, e.g., talker 2 /ma1/ vs. talker 1 /ma2/.

To control for an effect of Talker order, two randomised list of trials were created for each presentation mode (video and audio) and speech condition (normal vs. clear speech); half of the trials began with Talker 1 and the second half with Talker 2. Thus, there were two final versions of the experiment and the second version of the experiment was a counterbalanced version of version 1. Participants were randomly assigned to complete either version 1 or 2 of the experiment.

All participants completed a practice block before the beginning the experiment proper. There were 10 blocks in total. See Table (3) for a summary of the blocks. The order of blocks and trials were randomised across participants to control for any order effects.

Table 3

Summary of blocks

	Speech mode	
	Normal	Clear speech
Input mode		
AO	AO	AO
2D	AV	AV
	VO	VO
3D	AV	AV
	VO	VO

3. Results

A mixed analysis of variance (ANOVA) repeated measures model was used with language on 2 levels (Mandarin vs. AusE) as the between subjects factor and input mode on 5 levels (AO, 2D-AV, 3D-AV, 2D-VO and 3D-VO) and speech form with 2 levels (normal and clear speech) as within subjects factors. Mauchley's test of sphericity was violated for both the main effect of Input mode, $x^2 (9) = 20.68$, $p < .05$ and the interaction between language and Input mode, $x^2 (9) = 45.90$, $p < .05$, therefore degrees of freedom were corrected using the Huynh-Feldt estimates for the main effect ($\varepsilon = .90$) and interaction ($\varepsilon = .72$). The Huynh-Feldt correction was used for this comparison as the Greenhouse-Geisser value was greater than .75 and therefore too conservative (Fields, 2009).

There was a significant main effect of input mode on the percentage of correct responses, $F (3.60, 129.63) = 307.22$, $p < .001$. A series of pairwise comparisons, corrected using the Bonferroni adjustment for multiple comparisons, was carried out to further investigate the main effect of input mode. Comparisons between AO and 2D-AV and between AO and 3D-AV were not significant, $p > .05$, revealing no AV advantage. Further comparisons between 2D and 3D presentation effects on performance revealed no significant differences between 2D and 3D-AV conditions, $p > .05$ nor between 2D and 3D-VO conditions $p > .05$, suggesting no 3D advantage. Comparison between VO 2D and 3D conditions versus AO and AV 2D and 3D conditions revealed significantly lower percentages of correct responses in VO conditions. This indicates that participants were more accurate in identifying same/different tone contrast when they were provided with either auditory only or auditory and visual information, but not with visual information alone.

There was no significant main effect of speech mode on the percentage of correct responses, $F (1, 36)$, $p < .05$, but there was a significant main effect of language, $F (1, 36) = 85.77$, $p < .001$. Overall Mandarin listeners had a higher percentage of correct responses than

English listeners. However, there was no significant interaction between speech mode, and language, indicating that clear speech did not provide any advantage for both language groups.

There was a significant interaction between input mode and language, *F* (3.60, 129.63), *p* < .001. Looking at the interaction plot and means table revealed that Mandarin listeners performed better on AO and AV conditions than VO conditions. However, both language groups performed equally poorly on VO conditions. There was no significant three-way interaction between speech mode, input mode and language, indicating that there was no effect of speech mode on performance in any input mode and this is true for both language groups. Given these results, speech mode (normal vs. clear) was excluded as a factor from further analyses. See Table 4 for means and standard deviations of overall performance across input modes. See Figure 5 for the interaction plot.

Table 4

Means and standard deviation of percentage correct by normal/clear speech and language

	Mandarin				AusE			
	Normal		Clear		Normal		Clear	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AO	.98	.03	.98	.05	.75	.11	.79	.13
2D-AV	.98	.03	.99	.03	.76	.12	.76	.14
3D-AV	.98	.03	.98	.03	.77	.11	.76	.12
2D-VO	.52	.11	.52	.12	.52	.10	.47	.10
3D-VO	.48	.13	.51	.13	.51	.12	.52	.10

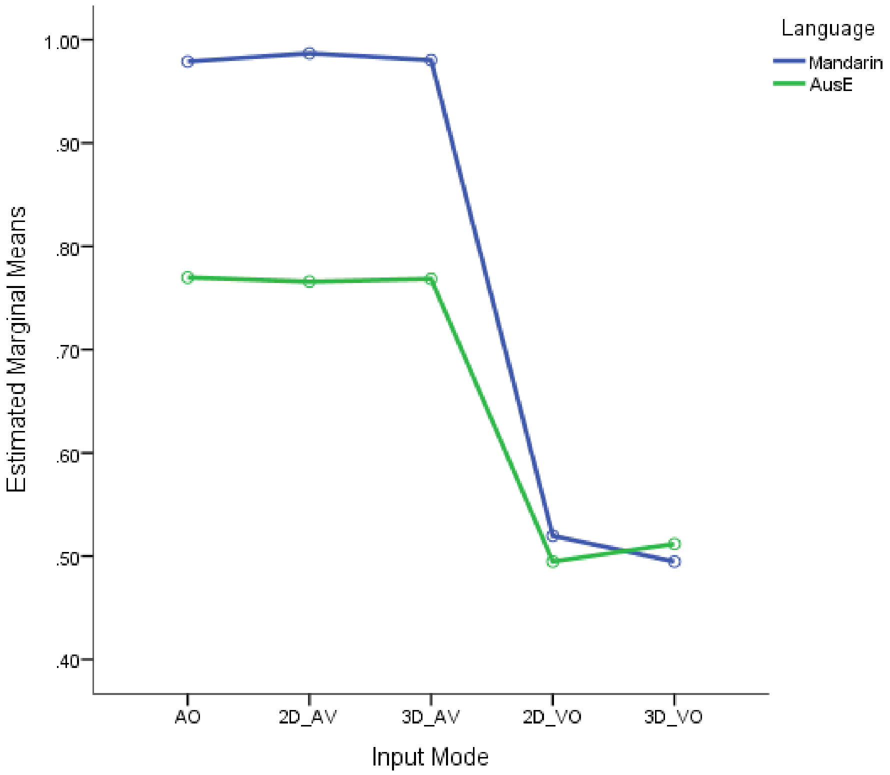

Figure 5 Interaction plot using estimated marginal means demonstrating interaction between input mode and language. Mandarin and AusE listeners are represented here using different coloured lines.

In order to investigate our exploratory hypothesis regarding the effect of visual salience of consonants on lexical tone discrimination, a separate repeated measures ANOVA was employed, with input mode and syllable on 2 levels (/ma/ and /na/) as within subjects factors and language as a between subjects factor. Mauchley's test of sphericity was violated for the interaction between input mode and syllable x^2 (9) = 41.62, *p* < .05, therefore degrees of freedom were corrected using the Greenhouse-Geisser estimates (ε = .68). Greenhouse-Geisser was used in this comparison as its value was less than .75 and therefore not too conservative (Fields, 2009).

There was a significant main effect of syllable, *F* (1, 36) = 15.09 *p* < .001. There was also a significant interaction between syllable and input mode, *F* (2.07, 98.01) = 5.00, *p* < .05. A series of within subjects contrasts were carried out to further investigate this interaction revealing a significant difference between 2D-VO and 3D-VO conditions and syllable, *F* (1, 36) = 5.66, $p < .05$. For both language groups, /ma/ contrasts had higher percentages of correct responses than /na/ contrasts and performance in 2D-VO conditions was better than 3D-VO conditions. See Table 5 for means and standard deviations of /ma/ and /na/ contrasts.

Table 5

Means and standard deviation of percentage correct for /ma/ and /na/ contrasts by Input mode and language

	Mandarin				AusE			
	/ma/		/na/		/ma/		/na/	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AO	.97	.05	.98	.03	.79	.09	.75	.12
2D-AV	.98	.30	.99	.02	.78	.13	.76	.12
3D-AV	.97	.04	.98	.03	.78	.10	.76	.10
2D-VO	.57	.13	.46	.10	.53	.09	.45	.10
3D-VO	.51	.11	.49	.10	.52	.09	.51	.10

Previous studies have indicated that visual information varied considerably across talkers (Dohen, Lœvenebruck & Harold, 2006; Chen & Massaro, 2008). Therefore it is possible that identifying unique visual information for each tone is difficult across talkers. As there may be less within tone variation across talkers compared to between tone variations, i.e., same vs. different tones, further analysis was conducted to address this. In order to investigate whether the percentage of correct responses varied across same (e.g. tone1 vs. tone1) versus different contrasts (e.g. tone1 vs. tone3), a mixed model repeated measures ANOVA was performed with same/different contrast included as an additional factor. There was no significant main effect of same/different contrasts on the percentage of correct responses. AusE listeners generally performed better on the same trials than the different trials however this comparison did not reach significance. See Table 6 for means and standard deviations of same and different contrasts.

Table 6

Means and standard deviation of percentage correct for same/different contrasts by input mode and language

	Mandarin				AusE			
	Same		Different		Same		Different	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AO	.98	.04	.98	.04	.82	.16	.74	.12
2D-AV	.98	*.05*	.99	.02	.83	.13	.73	.13
3D-AV	.97	.04	.98	.03	.78	.13	.76	.14
2D-VO	.52	.14	.51	.14	.50	.14	.50	.10
3D-VO	.52	.18	.49	.11	.50	.11	.52	.09

4. Discussion

In the present study we aimed to investigate whether 3D presentation could facilitate the perception of Mandarin lexical tone. Firstly, based on previous studies we predicted that listeners would be more accurate in tone discrimination performance for AV conditions compared to both AO and VO. This is because AV provides both auditory and visual information and is the most informative of all conditions. We further predicted that 3D presentation would provide additional useful visual information for listeners, therefore tone discrimination performance should be better in 3D than 2D conditions. As a secondary aim we sought to investigate whether exaggerated auditory information in clear speech was associated with more useful visual information. We therefore predicted better tone discrimination performance in clear speech compared to normal speech conditions. Additionally, we expected that native listeners (Mandarin) who are experienced tone language listeners would show better performance compared to inexperienced listeners (AusE). As a final exploratory hypothesis, we sought to investigate whether the visual salience of the syllable, visually salient /ma/ vs. ambiguous /na/, would influence the perception of lexical tone. We expected that presentations containing more visual information, both consonant and tonal, would facilitate visual speech perception so performance on /ma/ should be better than /na/ in AV and VO conditions.

Contrary to previous findings, we did not find any difference on accuracy in tone discrimination performance between AV and AO conditions. Mandarin listeners performed slightly better on the AV conditions but again this was not significantly different due to a ceiling effect. The significant main effect of AV input mode was, therefore, driven by the poor performance on VO conditions. However, both groups performed around chance on VO conditions suggesting visual information alone is not enough for tone discrimination. These results are somewhat different from previous research. In general, the AV literature on consonants, vowels and prosodic contrasts has found improvement in performance in AV conditions beyond that in AO and VO conditions (Summerfield & McGrath, 1984; Wang, Behne & Jiang, 2009; Cjevic, Kim & Davis, 2010). However, our findings are consistent with studies of AV tone perception which used similar discrimination paradigms. These studies also reported equal performance in AV and AO conditions in quiet listening conditions (Burnham, Ciocca & Stokes, 2001; Smith & Burnham, 2012). The absence of AV advantage indicates that listeners placed greater weight on auditory information when perceiving lexical tones under quiet listening conditions.

In addition, tone discrimination performance was not better in 3D than 2D conditions. This suggests that any additional visual information, including depth cues, was not useful in tone discrimination. This is true for both native- and non-native listeners suggesting that both experienced and inexperienced listeners did not find 3D visual information useful in tone discrimination. Thus, our results suggest that there was neither an AV nor a 3D advantage for lexical tone perception.

The lack of 3D advantage over 2D in tone perception conditions does not appear to support production studies that have identified rigid head movements associated with tone production (Burnham, et al., 2006; Attina, Gibert, Vatikiotis-Bateson & Burnham, 2010). However, it is possible that while 3D presentation may provide better depth cues related to rigid head movements, the listeners are either not using this information or not finding this information useful for tone perception. Given that previous studies only found visual benefit in listening in noise conditions, it is possible that any 3D advantage might only be obvious under noisy listening conditions where the listener is forced to attend to visual information.

It is also possible that the addition of a second talker may have introduced more variability in the stimuli, making it a more difficult listening task for non-native listeners. It is possible that for native listeners, visual information is not useful in optimum listening conditions as indicated by the ceiling level performance, but that the task difficulty masked any visual affect for the non-native listeners. This is no trivial task. A listener in this study must learn both variation across tones and across talkers for the same tones. It is possible that this cannot be achieved within a single test session. Indeed other studies have reported considerable variation in visual information across talkers, including eye-brow and rigid head-movements for the production of contrastive focus in French (Dohen, Loevenebruck & Harold, 2006). Chen and Massaro (2008) also informally described variations across talkers in the amount of visual information that accompanies the production of Mandarin tones; with some talkers producing larger movements than others. Thus, while the rigid head movements that accompany lexical tones may be stable within individual talkers, the magnitude and/or patterns of these movements may vary across talkers, creating more confusion for inexperienced listeners when they attempt to generalise tone across talkers. Additionally, to avoid using visual information that is unrelated to tone (e.g., position of head at onset and offset of each trial) for tone perception, listeners always saw two different talkers as opposed to two different tokens from the same talker. This would have further increased the difficulty of the task.

These task characteristics may also have contributed to the poor performance in VO conditions found in this study, which contrast with previous studies that reported significantly higher than chance performance on VO conditions (Burnham, Ciocca & Stokes, 2001; Chen & Massaro, 2008; Smith & Burnham, 2012). In the present study, it appears that both experienced and inexperienced listeners were affected by tone production variations across talkers. The use of a single talker to create stimuli could have facilitated VO discrimination by reducing variation within tones such as duration. For Mandarin tones, durational cues can signal the identity of the tone, as some are consistently longer and shorter, e.g. the dipping tone is longest in duration whereas the falling tone is the shortest (Massaro, Cohen & Tseng, 1985). By using two talkers, the present study minimised the opportunity for native listeners to use these cues to distinguish between tones, as there was both inter- and intra-talker variation in the duration of the tonal stimuli. These results are also inconsistent with results from studies on prosody which demonstrate that listeners can use visual information to discriminate silent videos of intonation contrasts (Cjevic, Kim & Davis, 2010). However, intonation contrasts require listeners to monitor head movement at the sentence level rather than at the syllable level as in the present study. It could be that a lack of familiarity with the use of pitch at the syllable level made AusE listeners less sensitive to visual tone information.

In terms of any facilitatory effect for tone discrimination using clear speech compared normal speech, we did not find any significant differences in performance for the two conditions. This was the case in all AV input modes, i.e., no difference between clear versus normal speech in AO, AV or VO conditions. This suggests that clear speech does not provide any additional useful information for tone discrimination over and above normal speech. While this study did not directly measure whether there are any differences in visual information between clear vs. normal speech, there were acoustic differences. These included longer durations and higher overall pitch across both talkers. Therefore, we can conclude that while acoustic adaptations were available, this did not assist listeners with better discrimination of tones. Perhaps for AusE listeners, who are unfamiliar with lexical tone, these additional variations hindered tone perception. While the influence of clear speech on tone perception has not been previously investigated, this finding differs from previous studies which reported improvements in intelligibility of words and sentences (Payton, Uchanski & Braida, 1994; Helfer, 1997). However, previous studies were conducted in varying levels of noise while this study was conducted in quiet only. Further study is needed to investigate if clear speech can facilitate lexical tone discrimination in adverse listening contexts.

The hypothesis that experienced native listeners would perform better was supported. Mandarin listeners had ceiling level performance on tone discrimination compared to AusE listeners in both AO and AV conditions. The significant interaction between language and Input mode, however, reflects the equally poor performance in VO conditions for both groups. Thus, in this study, language experience, while associated with improved performance in AO and AV conditions, was not associated with better performance when only visual information was available. These findings are entirely consistent with previous studies that found Mandarin listeners were more accurate than AusE listeners in tone discrimination but not in VO conditions (Burnham, Lau, Tam & Schoknecht, 2001; Smith & Burnham, 2012).

An interesting finding was that the visual salience of the syllable influenced lexical tone perception. Overall, better performance on tone discrimination was found for the bilabial /ma/ syllable when compared to the alveolar /na/ syllable. This effect was most prominent in the VO conditions, where performance was above chance for /ma/ contrasts and below chance for /na/ contrasts. This suggests that redundant visual information, consonant and tone in this case, does provide more useful information for tone discrimination. This has important implications. AV studies have demonstrated that the visual salience of consonants and vowels can influence the integration of auditory and visual information (Summerfield & McGrath, 1984; Wang, Behne & Jiang, 2009). Extending this notion to a tone discrimination task, raises the possibility that the initial visually salient consonant cued greater attention to the visual information for the syllable as a whole; eliciting greater integration between auditory and potential visual tone information. It is also possible that redundant visual information over the syllable, consonant and tone in this case, aids tone perception. Indeed other studies examining auditory only input have found that acoustic co-articulation cues are extremely useful for segmental and lexical perception (Scarborough, Styler & Zellou, 2011). It is, therefore, possible that a combination of redundant co-articulatory cues in both auditory and visual domains may help in lexical tone perception.

Taken together, our findings suggest that performance on AV tasks can be dependent on the paradigm used. The lack of AV advantage in the perception of lexical tone does not necessarily suggest an absence of functionally significant visual tone information. Rather, these results suggest the variability across talkers for tone production. This highlights the necessity for future studies to explore how talkers vary in the realisation of these movements and the magnitude of variation across talkers as well as how this might impact on lexical tone perception. Exploring AV perception in various listening conditions including in noisy conditions, which was not addressed here, will also help us to better understand the role of AV information on lexical tone perception. Also the finding that redundant auditory and visual information may be important for the learning of lexical tone warrants further investigation. This needs to be better addressed in future studies by using a range of syllables providing more opportunities for learning these co-articulations. This might also help with assessing within tone and across tone variations as well as intra-and inter-talker variations. Perhaps by addressing these issues in future designs we might see a 3D advantage in lexical tone perception.

In this preliminary study on 3D speech perception there are some limitations in the design. To manage participant fatigue, the length of the entire experiment was limited to 60 minutes. Logistically a lot of time was spent on changing between conditions and glasses between 2D and 3D presentations. We therefore chose only some conditions to explore here. While there was an exhaustive set of tone contrasts, all contrasts were presented in the same order, e.g. Rising vs. Dipping where rising was always presented first. However, there is no indication in the literature of directional effects for tone perception in Mandarin. See Francis & Ciocca (2003) for discussion of directional effects in Cantonese, which has a more complex tone system. Also the stimuli were only captured in citation form. This could have influenced the realisation of the pitch contour and duration of the word, plus it is devoid of acoustic tone co-articulations which would be present in connected speech. Another limitations is that the /na/ flat tone is not a meaningful word in Mandarin. While native raters did give this syllable a very high rating, 6 out of 6, in terms of naturalness, using a nonsense syllable may have influenced tone perception especially for the Mandarin listeners. Related to this is that for Mandarin listeners this is generally a lexical discrimination task but a nonsense syllable discrimination task for AusE listeners. The AusE listeners were therefore more likely to be using acoustic information. While this is often the case for most studies on AV speech perception, it is possible to control for this by using nonsense words for both groups. However, naturalness of tone productions was deemed more important for this study and for this reason mostly real words were selected as stimuli.

Beyond these limitations, this study opens many new avenues for future research. These results point to the importance of using more than one talker in the investigation of AV speech perception. Comparison between results from one talker paradigms and the multi-talker paradigm used in this study suggest that a single talker design may produce results that are not generalisable to other talkers. To assess whether auditory or visual information is useful, multiple talkers as well as multiple contexts (e.g., syllables) should be used. This point has important implications for how AV speech is represented and for language learning. It makes intuitive sense that exposure to a range of talkers and context is necessary for learning the natural variation and regularities that exists in speech production. Indeed this is how language is learned for most people.

This raises the question of how L2 learners are learning these variations and what their cognitive representations might be. The study of AV tone perception has been dominated by research comparing non-native vs. native performance. This neglects the group of L2 learners, who have been shown to be capable of using visual information to discriminate speech contrasts in their second language (Ortega-Llebaria, Faulkner & Hazan, 2001; Chen & Hazan, 2007). It would, therefore, be of interest to explore whether L2 learners of Mandarin, who have learned the meaningful distinction between tones, benefit from 3D-AV presentation of tone. Such a paradigm would also create more equality in task demands on participants. For non-native listeners, a tone discrimination task is acoustic by nature. In contrast, such a task would be more likely to evoke lexically based responses for L2 learners and native listeners of Mandarin.

In conclusion, the present study represents a first step in understanding the emerging potential of 3D technology for AV speech perception. The results also deepened our understanding of visual information for tone, by highlighting the important role of talker variation as well as acoustic and visual co-articulation for lexical tone perception. While this study did not find any benefits to 3D perception for lexical tone, this might not extend to other segments and suprasegments, nor to other languages with larger tone inventories such as Cantonese (6 tones) and Thai (5 tones). In language with a richer tone inventory the role of visual information might be more important for tone perception. There are also other possible applications for 3D technology, such as for hearing impaired populations who rely more on visual information for speech perception as well as for teaching a second language.

References

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,* (pp. 4-2).

Brooks, K. R., & Rafat, M. E. (2015). Simulation of driving in low-visibility conditions: Does stereopsis improve speed perception? *Perception abstract, 44*(2), 145-156.

Burnham, D., Ciocca, V., & Stokes, S. (2001). Auditory-visual perception of lexical tone. *Proceedings of Eurospeech Conference 2001, Aalsborg Denmark,* ISCA. Bonn, Germany, pp. 395-98.

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. In *Proceedings of International Conference on Auditory-Visual Speech Processing,* (pp. 155-160).

Burnham, D, Reynolds, J, Vatikiotis-Bateson, E, Yehia, H, Ciocca, V, Morris, R, Haszard, Hill, H, Vignali, G, Bollwerk, S, Tam, H & Jones, C. (2006). The perception and production of phones and tones: The role of rigid and non-rigid face and head motion, In *Proceedings of the 7th International Seminar on Speech Production,* (pp. 1-8).

Cvejic, E., Kim, J., & Davis, C. (2010). Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication, 52*(6), 555-564.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356-2366.

Chen, Y., & Hazan, V. (2007). Language effects on the degree of visual influence in audiovisual speech perception. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany,* (pp. 6-10).

Dees, T. M., Bradlow, A. R., Dhar, S., & (2007). Effects of noise on lexical tone perception by native and non-native listeners. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany,* (pp. 817-820).

Dohen, Lœvenbruck & Hill, 2006" (Dohen M., Lœvenbruck, H. & Hill, H., 2006. Visual Correlates of Prosodic Contrastive Focus in French: Description and Inter-Speaker Variability. In Proceedings of Speech Prosody 2006 (p. 221-224), Dresde (Allemagne), 2-5 May 2006).

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London, England: Sage.

Francis, A. L., & Ciocca, V. (2003). Stimulus presentation order and the perception of lexical tones in Cantonese. *The Journal of the Acoustical Society of America*, *114*(3), 1611-1621.

Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech, 21*(1), 1-33.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197-1208.

Hazan, V., & Kim, J. (2013). Acoustic and Visual Adaptations in Speech Produced to Counter Adverse Listening Conditions. In *Auditory-Visual Speech Processing (AVSP) 2013*.

Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 40*(2), 432-443.

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *The Journal of the Acoustical Society of America, 136*(3), 1352-1362.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*(8), 599-605.

Lam, A. K., Chau, A. S., Lam, W. Y., Leung, G. Y., & Man, B. S. (1996). Effect of naturally occurring visual acuity differences between two eyes in stereoacuity. *Ophthalmic and Physiological Optics, 16*(3), 189-195.

Lee, C. Y., Tao, L., & Bond, Z. S. (2010). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *Speech Communication, 52*(11), 900-910.

Massaro, D. W., Cohen, M. M., & Tseng, C. Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics,* 267-289.

McGurk H., MacDonald J. (1976). Hearing lips and seeing voices. *Nature, 264* (5588), 746–748.

Mixdorff, H., Charnvivit, P., & Burnham, D. K. (2005). Auditory-visual perception of syllabic tones in Thai. In *Proceedings of International Conference on Auditory-Visual Speech Processing,* (pp. 3-8).

Mixdorff, H., Hu, Y., & Burnham, D. (2005). Visual cues in Mandarin tone perception. In *INTERSPEECH* (pp. 405-408).

Mixdorff, H., Luong, M. C., Nguyen, D. T., & Burnham, D. (2006). Syllabic tone perception in Vietnamese. In *Proceedings of International Symposium on Tonal Aspects of Languages* (pp. 137-142).

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. In *Proceedings of International Conference on Auditory-Visual Speech Processing* (pp. 149-154).

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and hearing, 30*(6), 653-661.

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *95*(3), 1581-1592.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking Clearly for the Hard of Hearing IIAcoustic Characteristics of Clear and Conversational Speech. *Journal of Speech, Language, and Hearing Research*, *29*(4), 434-446.

Santoro, M., AlRegib, G., & Altunbasak, Y. (2012). Misalignment correction for depth estimation using stereoscopic 3-d cameras. In *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop,* (pp. 19-24).

Sato, M., & Howard, I. P. (2001). Effects of disparity–perspective cue conflict on depth contrast. *Vision Research, 41*(4), 415-426.

Scarborough, R., Styler, W., & Zellou, G. (2011). Nasal Coarticulation in Lexical Perception: The Role of Neighborhood-conditioned Variation. In *Proceedings of the 17th International Congress of Phonetic Sciences,* (pp. 1750-1753).

Sekiyama, K., & Tohkura, Y. I. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*(4), 427-444.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*(2), 420.

Smith, D., & Burnham, D. (2012). Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants). *The Journal of the Acoustical Society of America, 131*(2), 1480-1489.

Snowden, R., Thompson, P., & Troscianko. (2006). *Basic Vision: an introduction to visual perception*. Oxford, England: Oxford University Press Inc.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology, 36*(1), 51-74.

Takata, Y., & Náběłek, A. K. (1990). English consonant recognition in noise and in reverberation by Japanese and American listeners. *The Journal of the Acoustical Society of America, 88*(2), 663-666.

Tuntibundhit, C., et al. (2014). Perception of Thai Tones in Hearing Impaired and Cochlear Implant Patients. In *International Conference of Thai Studies.*

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181-1186.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America, 124*(3), 1716-1726.

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics, 37*(3), 344-356.

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain and Language*, *78*(3), 332-348.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones: Transfer to production. *The Journal of the Acoustical Society of America*, *106*(6), 3649-3658.

Ward, J. (2006). The Student's Guide to Cognitive Neuroscience. New York: Hove.

Wen, G., Markey, M. K., & Muralidlhar, G. S. (2014, March). A stereo matching model observer for stereoscopic viewing of 3D medical images. In *SPIE Medical Imaging* (pp. 90370Z-90370Z).

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *The Journal of the Acoustical Society of America*, *75*(6), 1866-1878.